SIMILARITY DETECTION BASED

ON SEMANTIC DISTANCE

THESIS

Presented to the Graduate Council of
Texas State University-San Marcos
in Partial Fulfillment
of the Requirements

for the Degree

Master of SCIENCE

by

Kimberly E. Adams, B.A.

San Marcos, Texas
August 2007

ACKNOWLEDGEMENTS

I would like to thank all of the members of my committee for the many answers and thoughtful assistance. Without the encouragement, guidance, and support of my thesis advisor, Dr. East, this work would not have been possible. She has been an invaluable resource, I could not ask for more.

Thank you to my son, Simon, for keeping me in good spirits and listening to algorithms and code fragments instead of more traditional bedtime stories.

I owe my deepest gratitude to my mother. When I expressed my desire to return to school she not only encouraged me, but provided me with every means to make it possible. For helping me make my dream a reality, I dedicate this work to her.

This manuscript was submitted on June 18, 2007.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

SIMILARITY DETECTION BASED

ON SEMANTIC DISTANCE

by

Kimberly Elaine Adams, B.A.

Texas State University-San Marcos

August 2007

SUPERVISING PROFESSOR: DEBORAH EAST

Combining basic language processing methodologies with the conceptual framework of WordNet, semantic distance is used to measure the similarity between documents. Noun and verb word concepts are transformed into paths that represent their physical location within the WordNet hierarchy. Path prefixes are compared using three distinct algorithms–each investigates a particular type of semantic distance. The results suggest that similarity can be derived from a conceptual hierarchy. The key to finding similarity lies in its precise definition.

CHAPTER 1

INTRODUCTION

1.1 The Dysphasia Effect

Imagine suffering a brief dysphasic episode while enjoying your morning paper. How do you determine what you're reading? How do you make sense out of words that have no apparent meaning?

You may break the words down into pieces, identifying their prefix, root, and suffix. Perhaps the smaller segments of words will ring a bell?

Or you may start at the beginning and look each word up in a dictionary. But what if the dictionary's definition is just as unfamiliar as the word? What if the words used in the definition are as meaningless as the word itself?

Then you may use a thesaurus. Instead of looking for a definition, it may be more beneficial to discover other words that mean the same thing. However, finding out that a 'whichymabob' is synonymous with a 'thingamabob' does nothing to explain what a 'whichymabob' is.

A better strategy would be to determine to what category each word belongs. If 'whichymabob' is synonymous with a 'thingamabob' and both of these elements are parts of a 'jawhosywhazzit', then regardless of what 'whichymabob' means, you know that you have a part of a 'jawhosywhazzit'.

Categorization takes advantage of the 'is-part-of' relationship. This basic relationship organizes data and sets a framework for understanding. The overall idea is that meaning can be derived from subsuming concepts.

Computer programs are not unlike dysphasic people. They both have the ability to act intelligently, make deductions, and learn. But more, they both have no concept of what things mean or how things are related. Similarity depends on a basic understanding of how things are related.

The focus of this thesis is to provide a semantic framework to allow a computer program to *understand* a document. Documents are transformed from sets of words into semantic networks. These networks represent a summation of the possible meanings that words in the document may have. Each path in the network is composed of a hierarchy of concepts, from the most general topic to a specific instance. Comparing one document's paths to the network of another, allows the degree of similarity to be evaluated. Similarity becomes a measurement of the semantic distance between two documents.

1.2 Similar, Not the Same

The process of deciding similarity is notably different from a matching problem. Where matching looks for the existence of element x in set A and in set B, similarity asks is set A like set B?

Matching depends on the symmetrical equivalence relation. It is identified through equality–a relation that evaluates to either true or false. Similarity is achieved through the comparison and intersection of the features of objects. Similarity is a relation that measures the degree to which things match.

1.3 Imitating Life

There are many ways the human mind assesses similarity. The most obvious method is conceptual distance. Conceptual distance implies a measurement based on conceptual proximity, but in fact it's a measurement of shared features. Objects that are conceptually closer together share a greater number of features.

The concept of chair is more similar to lumber than it is to book. A chair and a piece of lumber share a number of features: they are both solid; they're both made of wood; and they may even be used in similar functions like burn(x) or sitOn(x).

The differences between books and chairs are greater than the similarities. They're conceptually further apart and share few features. The only ties between them may be their co-occurrence in particular environments or the inappropriate use of one for the other. The lack of common features increases the conceptual distance.

There is a conceptual hierarchy that we innately understand. This hierarchy allows for us to understand the relationship between concepts. Lumber is understood to be a more elementary structure from which a chair (or house) is made. The fact that a chair is composed of lumber demonstrates inherent similarity.

This idea is also exemplified by the concept of family resemblance. This basic notion refers to features that are inherited from one generation to the next and are shared among family members. This is a prime example of how the 'is-part-of' relation is strongly tied to similarity assessment.

Without our conceptual framework, we are unable to assess similarity or determine meaning from symbols or objects. Without our framework, we are… dysphasic.

1.4 Overview

The introduction is designed to present the challenges of similarity assessment to the reader. From the introduction this paper launches into a presentation of research.

Chapter two defines the properties of similarity and describes methodologies for assessing similarity. The processes presented in the second chapter highlight the work of cognitive scientist, Amos Tversky. Tversky's research presents insights into how the human mind decides similarity and provides methods for measuring similarity. These findings have been applied to the problem of determining document similarity.

The assessment of similarity in this paper is focused on natural language processing. The goal specifically is to measure similarity between documents. Chapter three introduces the Reuter-21578 document data set. All documents that have been selected for similarity comparison come from this data set.

The forth chapter presents research regarding morphology. The topic of particular interest concerns the decomposition of words into stems through letter successor variety.

The fifth chapter describes Eric Brill's part-of-speech tagger. The tagging used during similarity assessment is directly derived from Brill's tagging algorithm. Tagging is an instrumental method of reducing ambiguity during semantic analysis.

Chapter six presents WordNet. This is the semantic framework from which meaning is derived and conceptual similarity assessed. There are three types of words considered in WordNet: nouns, verbs and modifiers. The hierarchical structure for each type of word in WordNet is unique. Each type is defined and its hierarchy described.

Chapter seven introduces the process created for similarity assessment. This chapter marks the beginning of unique content developed specifically for this thesis. Each step in similarity comparison is outlined and described.

Chapter eight gives a detailed explanation of the data structures and algorithms used to implement the similarity detection application.

Chapter nine provides result analysis and a mini-example detailing each step of the process. Modifications made to improve the functionality of algorithms are described and potential directions for future development are discussed.

Chapter ten presents conclusions drawn from the data presented and summarizes this accomplishment.

CHAPTER 2

SIMILARITY

Similarity is the basic relation that individuals use to generalize, classify and conceptualize. It is a complex and abstract concept. The very definition of similarity is often set by the context in which it occurs. In the broadest sense, similarity can be defined as a relation that measures the extent to which two objects are alike (and different).

Similarity relations are reflexive, asymmetric and intransitive. Equivalence relations are reflexive, symmetric and transitive. From this, it follows that similarity does not imply equivalence.

2.1 Similarity is Asymmetrical

A similarity statement is composed of a subject *a* and a referent *b*. Such statements are directional in nature, consider: "*a* is like *b*." The truth of this statement does not guarantee the truth of the converse, "*b* is like *a*." That is, the similarity relation "*a* is like *b*" is not equivalent to "*b* is like *a*".

The subject and referent are decided based on the relative salience of the objects. The more relevant term becomes the referent or prototype and the less notable becomes the subject or variant. "The direction of asymmetry is determined by the relative salience of the stimuli; the variant is more similar to the prototype than vice versa" [11, 328].

Consider the simile: “Your future is like a primed canvas.” In this simile, the subject is ‘your future’ and the prototype is ‘primed canvas’. This simile means that you can ‘paint’ your future anyway you choose. Now consider the converse, “A primed canvas is like your future.” This implies that your future looks rather blank. While either statement could be true, the semantics set by the direction are very different. Clearly, the similarity relation is not symmetrical.

2.2 Similarity is Not Transitive

Using the concept of triangle inequality, Tversky supposed that perhaps “if *a* is quite similar to *b*, and *b* is quite similar to *c*, then *a* and *c* cannot be very dissimilar from each other” [11, 329]. He illustrated the following example showing that the triangle inequality simply does not hold for similarity.

> “Consider the similarity between countries: Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of their political affinity); but Jamaica and Russia are not similar at all” [11, 329].

While the political alignment and names of countries change over time, this observation of intransitivity holds.

2.3 Features

Features are the set of attributes that define an object. They not only describe an object in terms of simple qualities, such as color, shape, or size–but also express abstract concepts like complexity or spatial configuration. Features should represent distinguishing characteristics, as features shared by all objects lack diagnostic value. Similarity is based on the comparison and distinction of feature sets. It measures not only the degree to which two objects possess matching features, but also the degree to which they differ.

The most important characteristic of a feature is its *salience*–the ability to stand out from its neighbors. Salience distinguishes prototypical characteristics. It has two components: diagnostic value and intensity.

A feature's diagnostic value is determined from the types of classifications that can be determined from it [11, 344]. A feature that sets a clear distinction for a set of objects will have a higher diagnostic value than one that only partially differentiates a set. For example, the diagnostic value of the feature 'warm-blooded' to the set of primates is insignificant, as all primates are warm-blooded. However, it will have significant value if the context is the set of all living creatures.

The intensity of a feature refers to the perceptual factors that remain relatively constant across different contexts and increase the signal-to-noise ratio. Intensity describes features like brightness, loudness, and size. Intensity refers to a range of values that a feature has which can be used to distinguish objects.

2.4 Categorization

There is a strong correlation between the features of an object and the class to which it belongs. The grouping of similar features into clusters forms basic categories. Categories reduce information load through the organization of features. Clusters at each level of abstraction should "maximize the similarity of objects within a cluster and the dissimilarity of objects from different clusters" [12, 91].

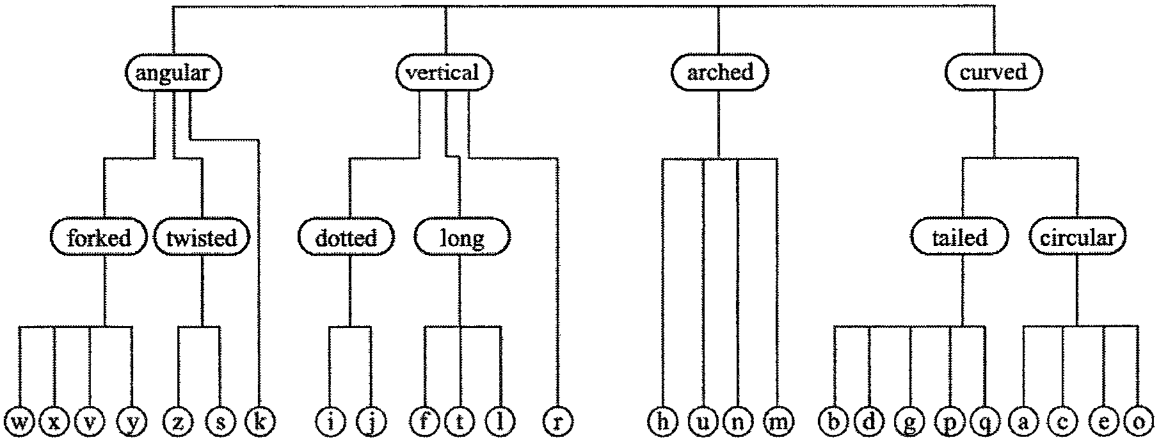


Figure 2.1 Representation of Letter Similarity as an Additive Feature Tree.

Figure 2.1 demonstrates how the similarity of letter forms can be represented through the clustering of features [11, 346]. The letters that share the most features are in fact, the most similar. Category resemblance, or the features that elements in a category share, is a measure of similarity.

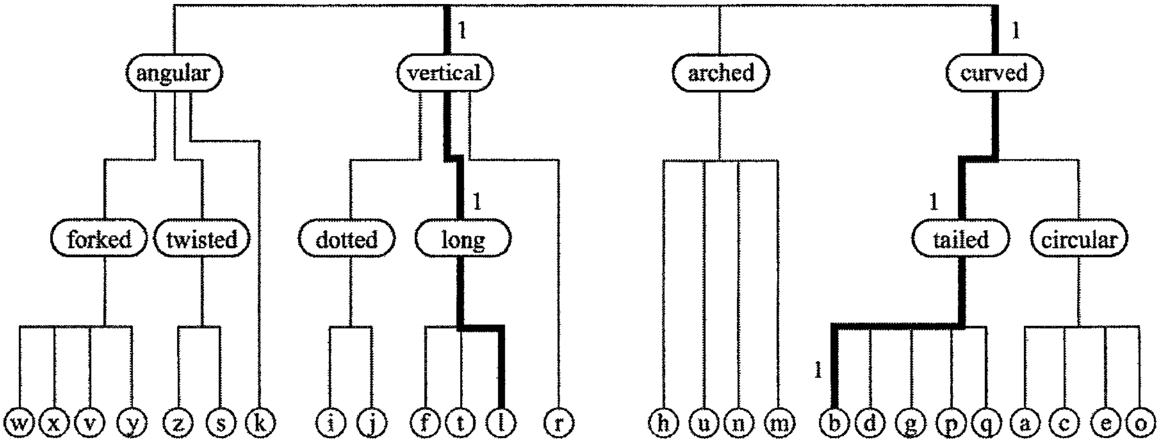


Figure 2.2 Distance Measure to Gauge Conceptual Distance.

Dissimilarity can be represented as a measurement of the distance between objects in different categories. As illustrated in Figure 2.2, the distance between the letter *l* and the letter *b* is the sum of the distances between them. If each arc has a distance of 1, then the distance between *b* and *l* is 6. The conceptual distance between any two letters is directly related to how many features they share and how similar they are.

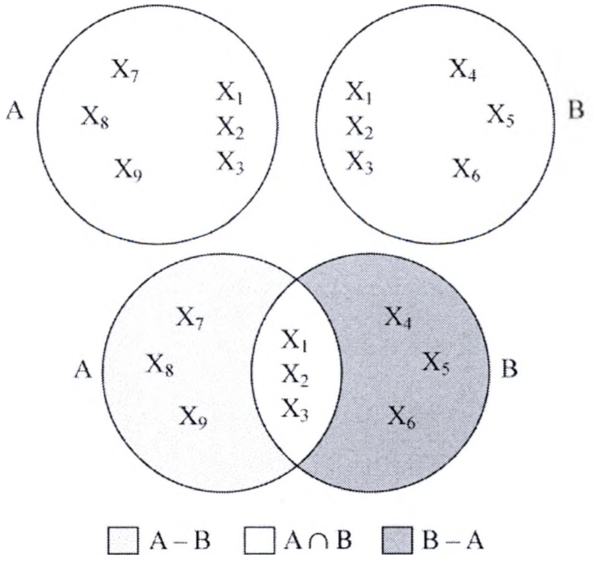
2.6 Axiomatic Theory of Similarity

The Axiomatic Theory of Similarity defines similarity in terms of set theory. For these five axioms, the domain is the set of objects {*a*, *b*, *c*} and A, B, and C represent the sets of features for objects *a*, *b*, and *c*.

The first axiom, *Matching,* explains that the measure of similarity between objects is increased by the co-occurrence of features and diminished by the difference.

"Axiom 1. *Matching*: s(*a*, *b*) = F(A ∩ B, A – B, B – A)

The similarity of *a* to *b* is expressed as a function F of three arguments: A ∩ B, the features that are common to both *a* and *b*; A – B, the features that belong to *a* but not to *b*; and B – A, the features that belong to *b* but not to *a*" [11, 330].



$s(a, b) = F(\{X_1, X_2, X_3\}, \{X_4, X_5, X_6\}, \{X_7, X_8, X_9\})$

Figure 2.3 Axiom 1: Matching.

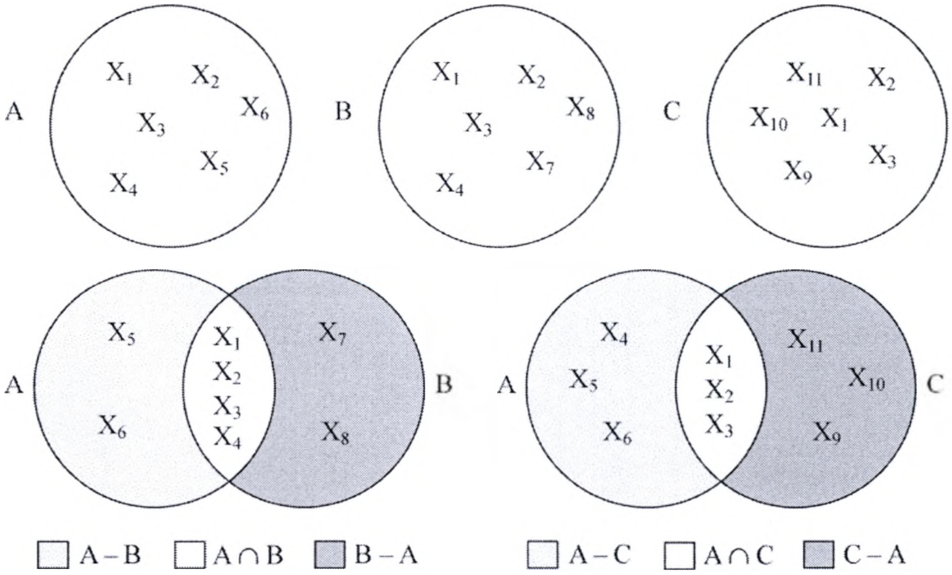
The crux of this axiom is that matching in similarity considers *all* of the features in A and B, not only the features that A and B share. Figure 2.3 demonstrates this point visually. The inclusion of the differences between sets distinguishes this type of matching as a similarity process.

The second axiom, *Monotonicity,* describes how similarity increases and decreases.

"Axiom 2. *Monotonicity*: $s(a, b) \geq s(a, c)$ whenever $A \cap B \supset A \cap C$,

$A - B \subset A - C$, and $B - A \subset C - A$.

Similarity increases with addition of common features and/or deletion of distinctive features" [11, 330]. Figure 2.4 shows an example of how the similarity of A to B is greater than the similarity of A to C, if A and B have more common elements and fewer distinct elements than A and C.



$$A \cap B = \{ X_1, X_2, X_3, X_4 \} \supset A \cap C = \{ X_1, X_2, X_3 \}$$
$$A - B = \{ X_5, X_6 \} \subset A - C = \{ X_4, X_5, X_6 \}$$
$$B - A = \{ X_7, X_8 \} \subset C - A = \{ X_9, X_{10}, X_{11} \}$$

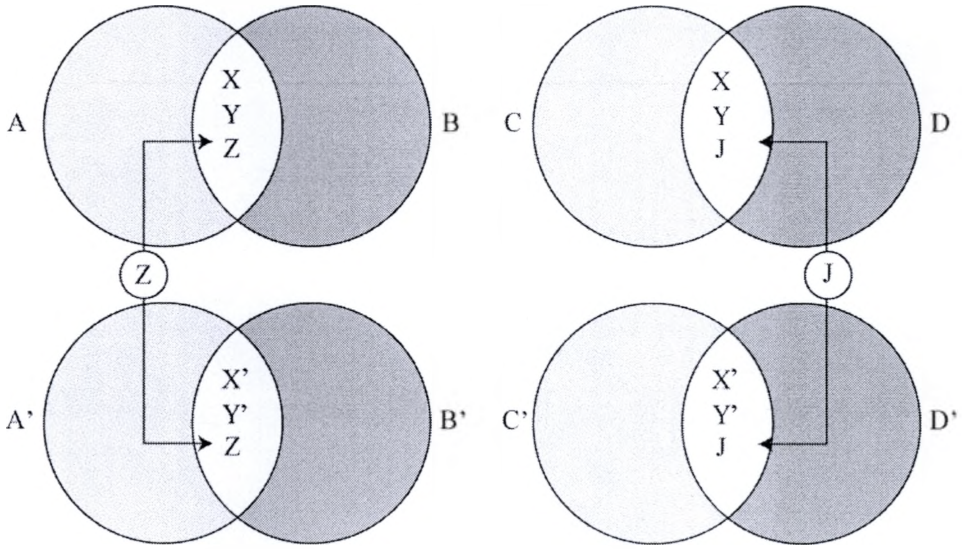
Figure 2.4 Axiom 2: Monotonocity.

Independence demonstrates that features are not conditionally bound to each other. The occurrence of one event does not affect the occurrence of another.

"Axiom 3. *Independence*: Suppose pairs (*a*, *b*) and (*c*, *d*) as well as pairs (*a*', *b*') and (*c*', *d*') agree on the same two components, while the pairs (*a*, *b*) and (*a*', *b*') as well as the pairs (*c*, *d*) and (*c*', *d*') agree on the remaining (third) component. Then

s(*a*, *b*) ≥ s(*a*', *b*') iff s(*c*, *d*) ≥ s(*c*', *d*').

Thus, the ordering of the joint effect of any two components is independent of the fixed level of the third factor." [11, 331]. Figure 2.5 demonstrates this concept.



{ X, Y, Z } ≥ { X' Y' Z } iff { X, Y, J } ≥ { X' Y' J }

Figure 2.5 Axiom 3: Independence.

Solvability, requires feature sets to be robust enough to allow for similarity evaluation. Objects must be described thoroughly and each feature must be essential. Over generation of features can lead to redundancy, while too few features may fail to uncover appropriate matches.

The last axiom, *Invariance,* requires that the range in feature values be consistent across all objects. This ensures that the values associated with the features in one object are identical to the values associated with the features in a different object. The comparison of like terms is necessary for similarity detection.

2.7 Contrast Model

Any function that fulfills the requirements of the first two axioms is considered to be a *matching function.* The most basic matching function Tversky describes is the Contrast Model.

The Representation Theorem, assuming the five assumptions from the Axiomatic Theory of Similarity hold, states: "Then there exists a similarity scale S and nonnegative scale f such that for a, b, c, and d in Δ, where Δ is the set of all objects being considered.

(i). $S(a, b) \geq S(c, d)$ iff $s(a, b) \geq s(c, d)$;

(ii). $S(a, b) = \Theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$, for some $\Theta, \alpha, \beta \geq 0$

(iii). f and S are interval scales" [11, 332].

From the contrast model, Tversky establishes that similarity is both a measure the similarity and the difference of feature sets.

2.7.1 Measuring Conceptual Distance

Figure 2.6 shows that the Contrast Model can be used to measure conceptual distance. If α and β equal zero and Θ equals one, then the similarity of a and b will be the derived from the common features of a and b.

$$\Theta = 1, \alpha = 0, \beta = 0$$
$$S(a, b) = 1 \cdot f(A \cap B) - 0 \cdot f(A - B) - 0 \cdot f(B - A)$$
$$S(a, b) = f(A \cap B)$$

Figure 2.6 Conceptual Similarity.

Now suppose that α and β equal one and Θ equals zero, then the similarity of *a* and *b* will be a measurement of the dissimilarity of *a* and *b*. The pure conceptual distance between *a* and *b* is demonstrated in Figure 2.7.

$$\Theta = 0, \alpha = 1, \beta = 1$$
$$S(a, b) = 0 \cdot f(A \cap B) - 1 \cdot f(A - B) - 1 \cdot f(B - A)$$
$$S(a, b) = f(A - B) - f(B - A)$$

Figure 2.7 Conceptual Distance.

The equations of the Contrast Model demonstrate that similarity is not symmetrical:

"$s(a, b) = s(b, a)$ iff $\alpha f(A - B) + \beta f(B - A) = \beta f(A - B) + \alpha f(B - A)$

iff $(\alpha - \beta) f(A - B) = (\alpha - \beta) f(B - A)$.

Hence, $s(a, b) = s(b, a)$, if either $\alpha = \beta$ or $f(A - B) = f(B - A)$" [Tversky, 333].

From this, it's evident that similarity is asymmetric, as symmetry is only supported when the comparisons are not directional ($\alpha = \beta$) or the feature sets that describe the objects are identical ($f(A - B) = f(B - A)$).

The value of the intersection between two sets of features is calculated by the following equation, presented by Tversky. "Let X_a denote the proportion of subjects who attributed feature X to object *a*, and let N_x denote the number of objects that share feature X. For any a, b, define the measure of their common features by $f(A \cap B) = f(A \cap B) - \sum X_a X_b / N_x$, where the summation is over all X in $A \cap B$, and the measure of their distinctive features by $f(A - B) + f(B - A) = \sum Y_a + \sum Z_b$ where the summations range over all $Y \in A - B$ and $Z \in B - A$, that is, the distinctive features of a and b, respectively" [11, 338].

2.8 Ratio Model

Another method of assessing similarity is the Ratio Model. This method measures similarity by measuring the proportion of sameness less the proportion of difference. The value of similarity is normalized to fall between 0 and 1, Figure 2.8. This shows the degree of similarity between two sets of features.

$$s(a,b) = \frac{f(a \cap b)}{f(a \cap b) + \alpha f(A - B) + \beta f(B - A)}, \alpha\beta \geq 0.$$

Figure 2.7 Ratio Model.

Normalized values can be fairly compared no matter how different their feature space. The values for α and β serve a variety of purposes. A higher α value than β can be used to force directionality in the comparison. Setting either value to zero will cause the ratio to be a straight proportion of the other. Setting both values to one results in the ratio of $A \cap B / A \cup B$.

2.9 Product Moment Model

When studying how humans determine similarity, Tversky conducted a number of studies. One study used twelve types of vehicles as input, and asked the participants to quickly list the features that described each of the twelve vehicles. The resulting feature lists were then compared to each other. The subjects created 324 features, of which 100 were shared between two or more vehicles. A similarity score for every pair of vehicles was determined based on a product-moment correlation.

2.9.1 Pearson Product Moment Correlation Coefficient

The idea of measuring the similarity of objects may be based on the product-moment led to the selection of the Pearson Product Moment Correlation Coefficient, Figure 2.8. This algorithm presents a powerful method for determining similarity based on feature comparison.

This equation looks at pairs of objects, the features that are shared by both objects are counted (X) and the features that apply to one object but not the other are counted (Y). The total number of features present in all objects are calculated. The mean is derived for features that co-occur and for those that are distinct. The X deviation and Y deviation for each object is assessed and the product is summed. This value is normalized by the product of the standard deviations of X and Y with the total number of objects.

$$r = \frac{\sum(X - \mu_X)(Y - \mu_Y)}{N\sigma_X\sigma_Y}$$

Figure 2.8 Pearson Product Moment Correlation Coefficient Formula.

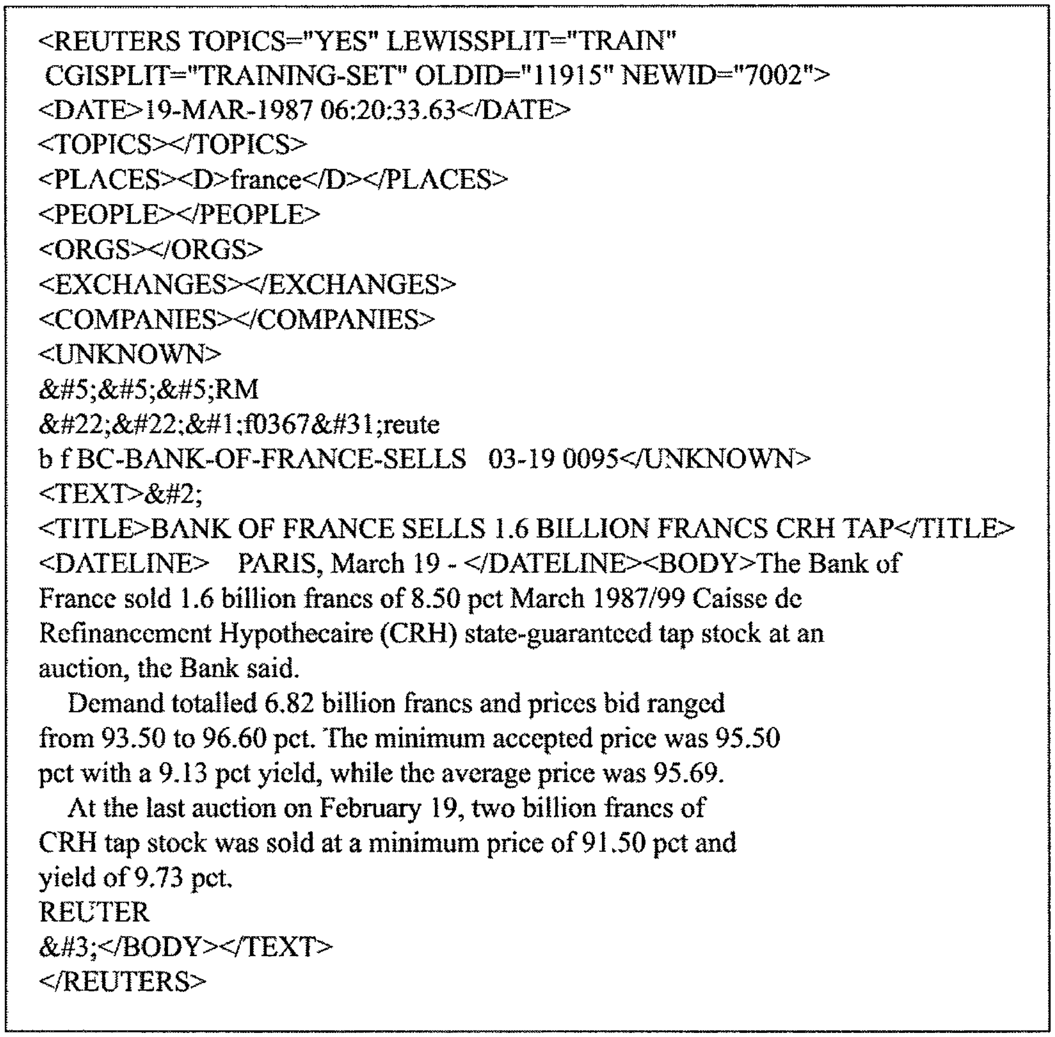
CHAPTER 3

THE DATA SET

The input used in evaluating this system is the set of documents, Reuters-21578, Distribution 1.0 [6]. Factors that contributed to the selectıon of this set of documents include the large number of formatted documents available, the variety of topic coverage and the randomness of order. Additionally, each document has been encoded with SGML tags. The SGML tags clearly define a document's boundary and offer abbreviated heading, title, date, and topic listings.

3.1 A New Use for Old News

The Reuters dataset consists of 21,578 SGML-tagged documents. Each document is a unique news story that has been preprocessed into SGML format and given a distinct identification number. Tagged data specifies the story's date, topics, places, people, organizations, exchanges, company, title, dateline, and body where data exists. Figure 3.1 displays the format and content of document number 7002–this is a random example from the Reuters-21578 collection. Of the 21,578 Reuters documents, 250 have been selected for similarity evaluation.



```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN"
 CGISPLIT="TRAINING-SET" OLDID="11915" NEWID="7002">
<DATE>19-MAR-1987 06:20:33.63</DATE>
<TOPICS></TOPICS>
<PLACES><D>france</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;RM
&#22;&#22;&#1;f0367&#31;reute
b f BC-BANK-OF-FRANCE-SELLS   03-19 0095</UNKNOWN>
<TEXT>&#2;
<TITLE>BANK OF FRANCE SELLS 1.6 BILLION FRANCS CRH TAP</TITLE>
<DATELINE>    PARIS, March 19 - </DATELINE><BODY>The Bank of
France sold 1.6 billion francs of 8.50 pct March 1987/99 Caisse de
Refinancement Hypothecaire (CRH) state-guaranteed tap stock at an
auction, the Bank said.
    Demand totalled 6.82 billion francs and prices bid ranged
from 93.50 to 96.60 pct. The minimum accepted price was 95.50
pct with a 9.13 pct yield, while the average price was 95.69.
    At the last auction on February 19, two billion francs of
CRH tap stock was sold at a minimum price of 91.50 pct and
yield of 9.73 pct.
REUTER
&#3;</BODY></TEXT>
</REUTERS>
```

Figure 3.1 Reuters SGML Format.

3.2 The Useful Subset

During evaluation, three types of documents were identified as problematic. These are documents that are either: too short; contain no sentences; or represent tabular data, intended to express horizontal information.

Short documents are defined as news articles that contain less than 50 words. These have been removed because they lack the feature depth necessary for meaningful similarity evaluation. Documents with no grammatical sentences have been removed as they contain data without context.

Documents that contain tabular data, that's meant to be read vertically, are removed because their context is lost when read row by row. When vertical column headings and entries are concatenated horizontally the process is corrupted. The result is the creation of ill-formed sentences that contain inappropriate designations of proper nouns and that fail to be tagged correctly. The errors ripple through the process of feature identification and evaluation, and cause unexpected results.

Beyond documents removed for problematic content–documents have been adjusted for consistency. Inconsistencies discovered during evaluation, have been corrected. These include: sentences that end without punctuation; sentences that end with multiple punctuation marks; inconsistent use of apostrophes and quotation marks; and incomplete SGML tags.

Consistent use of punctuation marks is required for accurate processing. These grammatical cues are expected to not only be present in each document but to follow standard rules. Document that don't contain consistent punctuation mark usage are modified so they: end each sentence in a single period; use apostrophes to indicate contractions and possession; use quotation marks to indicate with quotations; and each opening SGML tag has a matching closing SGML tag.

Worth noting, other errors, such as erroneous facts, misspelled words, etc., have been ignored. Documents are evaluated, for the most part, in their original state.

Chapter 4

MORPHOLOGY

4.1 Decomposition

Learning to read is a daunting task. The first requirement is a definition of the alphabet that attaches sounds to letter forms. Once letter forms are understood, the next step is to begin to sound out letters in words. After a few iterations, certain sounds that frequently occur together can be grouped into word fragments. With time and practice, the ability to distinguish the prefixes and suffixes of root words develops.

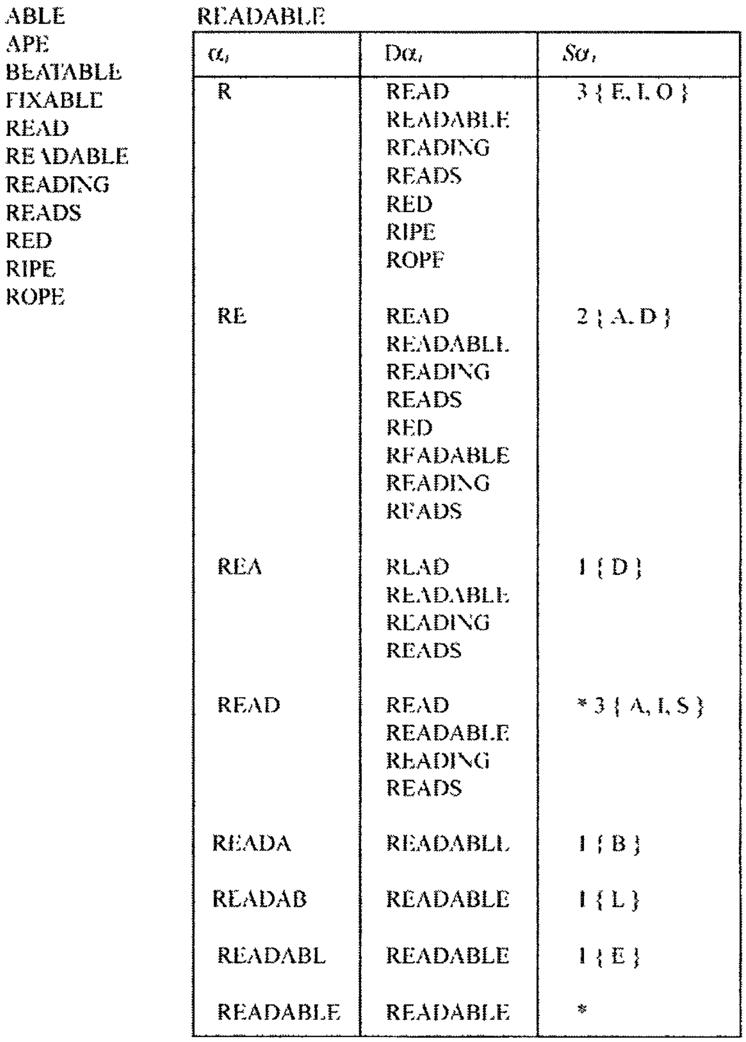
A similar phenomenon occurs with word stemming. A word is a sum of ıts parts. Identifying the pieces is a pivotal step to determining what the words mean.

4.2 Letter Successor Variety

Word segmentation is a process of decomposing words into well-defined parts. Letter successor variety relies on the statistical properties of the next letter in a word, given a body of text. This method works because in a robust corpus the structural properties of words themselves indicate where segmentation occurs.

Letter successor variety, as defined by Margaret Hafer and Stephen Weiss: "Let α be a word of length n; α_i is a length *i* prefix of α. Let D be a corpus of words. Dα_i is defined as the subset of D containing those words whose first i letters match α_i exactly. The successor variety of α_i, denoted $S\alpha_i$, is then defined as the number of distinct letters that occupy the i+1st position in words of $D\alpha_i$" [5, 372].

The big picture here is that the ith letter of a word depends on the (i – 1) letters that precede it. Specifically, letter successor variety identifies the number of letters that could come next. Another interesting property of letter successor variety is that it tends to decrease left to right, but increases at word boundaries.



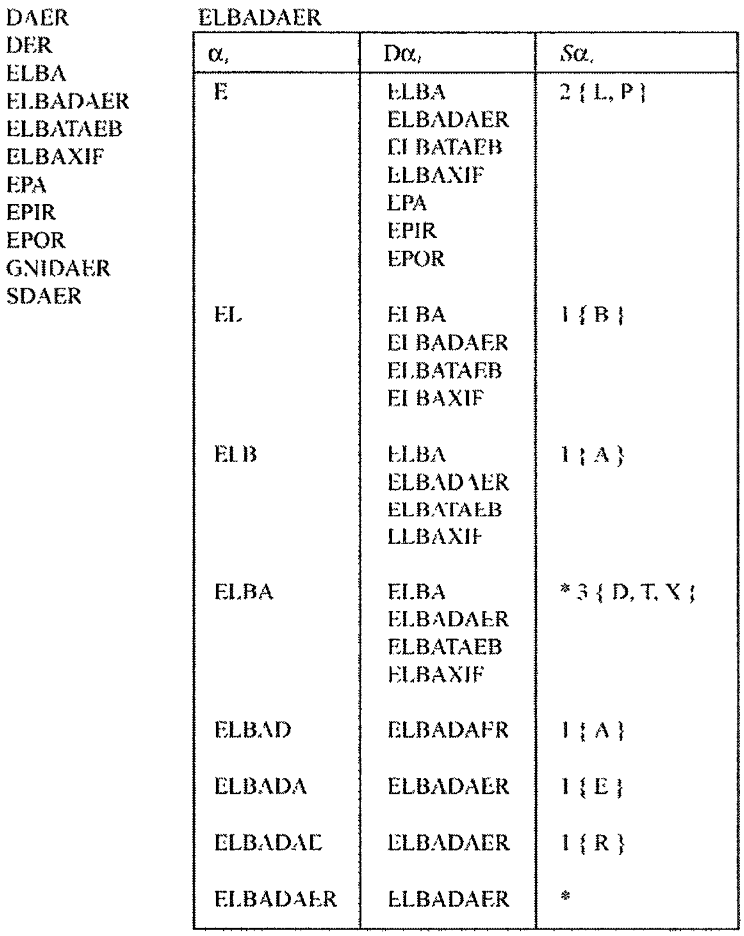
ABLE
APE
BEATABLE
FIXABLE
READ
READABLE
READING
READS
RED
RIPE
ROPE

READABLE

α_i	$D\alpha_i$	$S\alpha_i$
R	READ READABLE READING READS RED RIPE ROPE	3 { E, I, O }
RE	READ READABLE READING READS RED READABLE READING READS	2 { A, D }
REA	READ READABLE READING READS	1 { D }
READ	READ READABLE READING READS	* 3 { A, I, S }
READA	READABLE	1 { B }
READAB	READABLE	1 { L }
READABL	READABLE	1 { E }
READABLE	READABLE	*

Figure 4.1 Letter Successor Varieties for 'READABLE'.

Figure 4.1 shows an example of letter successor variety for the word "READABLE", given the context of the eleven words appearing to the left of the chart. When i is one, α_i is R. The words in the corpus that share the first i letters are: READ, READABLE, READING, READS, RED, RIPE, and ROPE. The next letters that could follow R are: E, I, and O. Because only three letters could come next, the successor variety of R is three. The chart demonstrates the letter successor variety for each successive letter in READABLE. An asterisk indicates the end of a word, which is useful in determining root/suffix barriers. Notice how successor variety decreases until an end word is encountered (i.e. when i equals four and α_i is READ).



DAER
DER
ELBA
ELBADAER
ELBATAEB
ELBAXIF
EPA
EPIR
EPOR
GNIDAER
SDAER

ELBADAER

α_i	$D\alpha_i$	$S\alpha_i$
E	ELBA ELBADAER ELBATAEB ELBAXIF EPA EPIR EPOR	2 { L, P }
EL	ELBA ELBADAER ELBATAEB ELBAXIF	1 { B }
ELB	ELBA ELBADAER ELBATAEB ELBAXIF	1 { A }
ELBA	ELBA ELBADAER ELBATAEB ELBAXIF	* 3 { D, T, X }
ELBAD	ELBADAER	1 { A }
ELBADA	ELBADAER	1 { E }
ELBADAE	ELBADAER	1 { R }
ELBADAER	ELBADAER	*

Figure 4.2 Letter Successor Varieties for 'ELBADAER'.

Segmentation is increased in accuracy by looking at words in reverse-order. This allows for the discernment of prefix/root barriers. Figure 4.2 demonstrates letter successor varieties for ELBADAER, given the same eleven words, this time considering them in reversed order. Again note the decrease in successor variety until an end of word is encountered (i.e. when *i* equals four and α_i is ELBA).

4.3 Entropy

Not only can word stems be identified through successor variety, once the framework is in place to store successor variety, that same structure can be used to examine the distribution of letters. *Entropy*–refers to the probability of finding a certain state based on the multiplicity of that state. "Using entropy in the segmentation process allows the importance of each successor letter to be weighted by its probability of occurrence [5, 375].

In a completely intelligent system, entropy is an encouraging method to deciding segmentation when no obvious delineation exists. Heiss defines entropy as: "Let $|D\alpha_i|$ be the number of words in the corpus that match an *i* letter prefix of the test word α. Let $|D\alpha_i|$ be the size of the subset of $D\alpha_i$ in which the $(i + 1)^{st}$ letter of the alphabet is $|D\alpha_{ij}| \setminus |D\alpha_i|$" [5, 375]. The entropy of the successor system for a test word prefix α_i is shown in Figure 4.3.

$$H\alpha_i = \sum_{j=1}^{26} \frac{D\alpha_{ij}}{D\alpha_i} \cdot \log_2 \frac{D\alpha_{ij}}{D\alpha_i}.$$

Figure 4.3 Entropy Equation.

An example of how entropy can be used in morpheme analysis is demonstrated in Figure 4.4. This string trie contains the words from the sentence: "To a top the town tore through the towns torrent of the tops." This sentence was created to force alliteration (and branching) in order to demonstrate letter successor variety and entropy. It is a purely syntactic example.

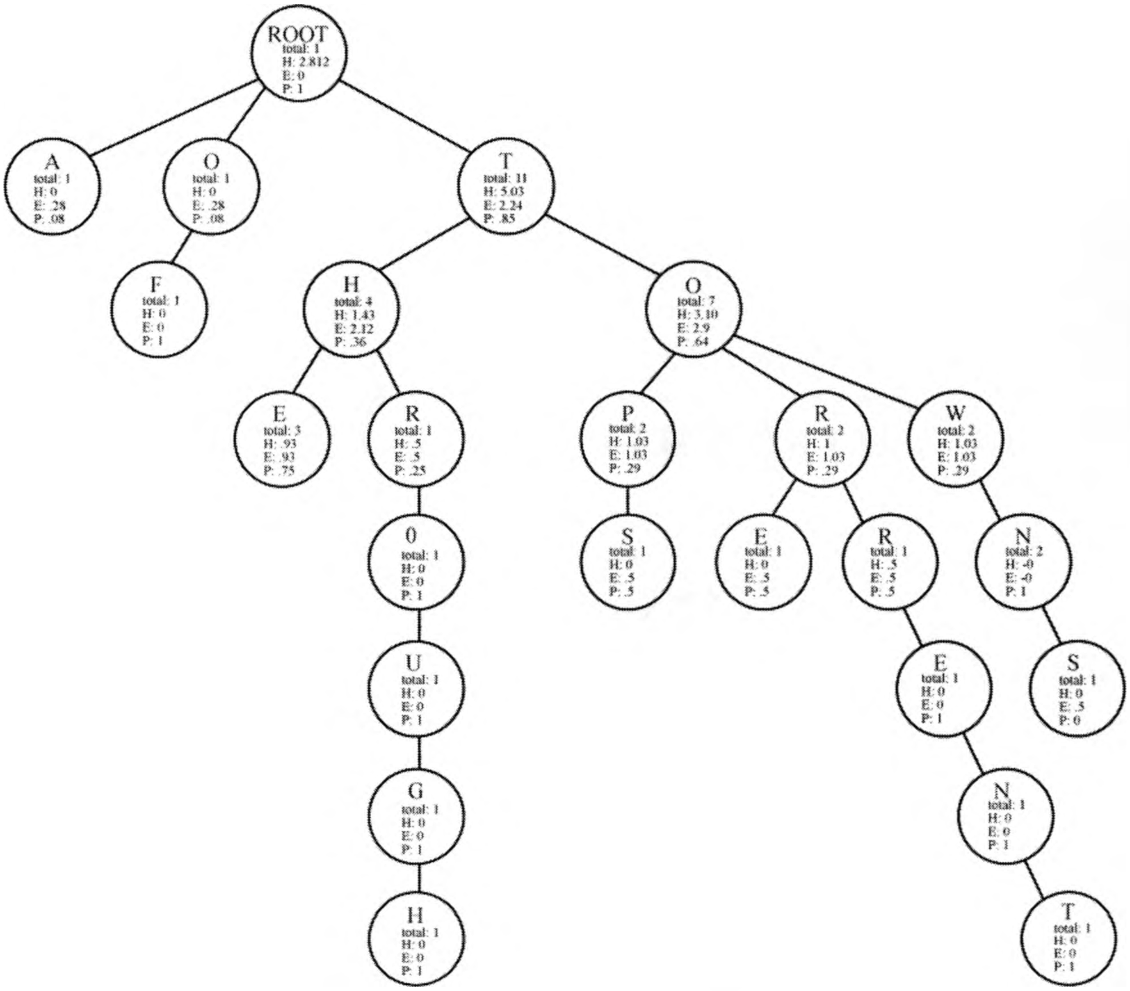


Figure 4.4 Entropy.

All of the words are read and stored as letter nodes in a string trie. In a depth first traversal, the counts of children of each node are passed up to the root. In a third pass, this time a breadth-first traversal, the entropy of each node is set in two steps. Each node's entropy is discovered using the formula -(L/C) $\log_2$ (L/C), where L is the total number of words ending at or beneath a node and C is the total number of words ending at or beneath a child node. The local entropy is stored in the variable E. The H value of each parent node consists of the summed values of all its children's E values. The probability (P value) of the children of each parent node sums to one. Each value P can be expressed as the ratio C:L.

Node Letter	Entropy	Child Total	Local Total	Parent Letter	Parent's Probability
A	0.285	1	13	ROOT	0.285
O	0.285	1	13	ROOT	0.569
T	0.408	11	13	ROOT	0.977
F	0	1	1	O	1
H	1.061	4	11	T	1.061
O	1.245	7	11	T	2.306
E	.311	3	4	H	.311
R	.5	1	4	H	.811
O	0	1	1	R	1
U	0	1	1	O	1
G	0	1	1	U	1
H	0	1	1	G	1
P	0.516	2	7	O	.516
R	1.03	2	7	O	1.549
W	.516	2	7	O	2.066
S	.5	1	2	P	1
E	.5	1	2	R	.5
R	.5	1	2	R	1
E	0	1	1	R	1
N	0	1	1	E	1
T	0	1	1	N	1
N	0	2	2	W	1
S	.5	1	2	N	1

Figure 4.5 Entropy Values.

Figure 4.5 reveals the order that values are set and summed. Row by row, each entry describes a node's evaluation in a breadth first traversal of the tree. The evaluation process sets the entropy of a node and sums it to its parent's *H* value. The field 'Node Letter' indicates the letters in the nodes of the tree in Figure 4.4. Note that entropy is directly affected by branching factor (successor variety) and frequency.

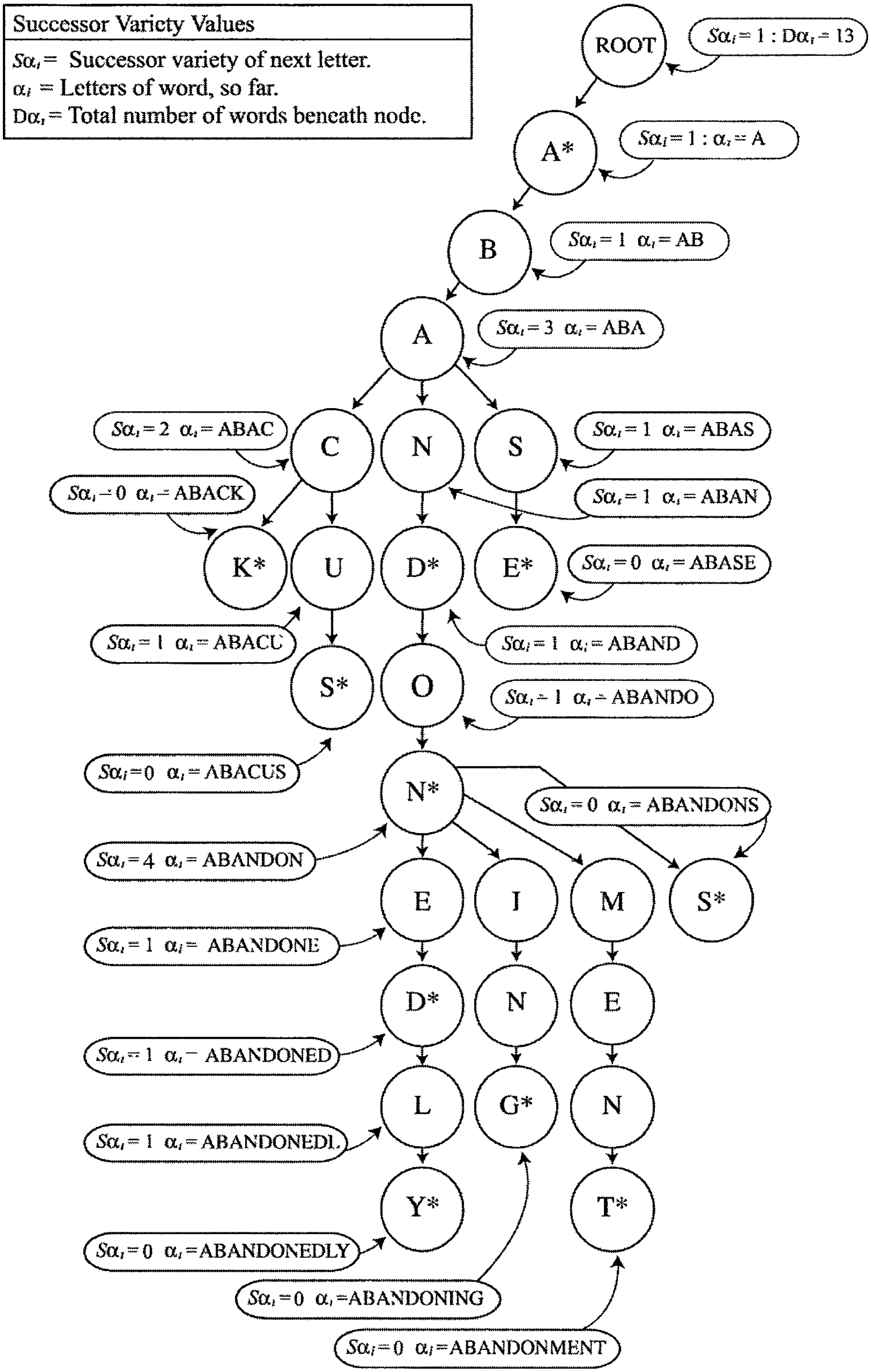

Successor Variety Values
$S\alpha_i$ = Successor variety of next letter.
α_i = Letters of word, so far.
$D\alpha_i$ = Total number of words beneath node.
ROOT
$S\alpha_i = 1 : D\alpha_i = 13$
A*
$S\alpha_i = 1 : \alpha_i = A$
B
$S\alpha_i = 1 \quad \alpha_i = AB$
A
$S\alpha_i = 3 \quad \alpha_i = ABA$
C
N
S
$S\alpha_i = 2 \quad \alpha_i = ABAC$
$S\alpha_i = 1 \quad \alpha_i = ABAS$
$S\alpha_i - 0 \quad \alpha_i - ABACK$
$S\alpha_i = 1 \quad \alpha_i = ABAN$
K*
U
D*
E*
$S\alpha_i = 0 \quad \alpha_i = ABASE$
$S\alpha_i = 1 \quad \alpha_i = ABACU$
$S\alpha_i = 1 \quad \alpha_i = ABAND$
S*
O
$S\alpha_i - 1 \quad \alpha_i - ABANDO$
$S\alpha_i = 0 \quad \alpha_i = ABACUS$
N*
$S\alpha_i = 0 \quad \alpha_i = ABANDONS$
$S\alpha_i = 4 \quad \alpha_i = ABANDON$
E
I
M
S*
$S\alpha_i = 1 \quad \alpha_i = ABANDONE$
D*
N
E
$S\alpha_i = 1 \quad \alpha_i - ABANDONED$
L
G*
N
$S\alpha_i = 1 \quad \alpha_i = ABANDONEDL$
Y*
T*
$S\alpha_i = 0 \quad \alpha_i = ABANDONEDLY$
$S\alpha_i = 0 \quad \alpha_i = ABANDONING$
$S\alpha_i = 0 \quad \alpha_i = ABANDONMENT$

Figure 4.6 Successor Variety.

4.4 Successor Variety in String Tries

Clear identification of where stemming occurs is useful in preparing words for semantic analysis. Working with root words, rather than root + tense words simplifies lexical queries and increases the odds of finding search words.

Figure 4.6 shows how successor variety is reflected in a string trie structure. Successor variety is the number of children of each node. When the number of children peaks, a word stem has been identified. The strmg trie does structurally what successor variety does algorithmically.

CHAPTER 5

PART OF SPEECH TAGGING

Language is ambiguous. Where spoken language is confused with *homophones*–words that sound the same but have different meanings, written language is riddled with *heteronyms*–words that are spelled the same but have multiple meanings.

Ambiguity is a huge problem that leads to the improper categorization of features. The words 'plant cotton' and 'cotton plant' imply two different concepts. The first fragment suggests an action–in this case 'plant' is a verb. The second fragment describes an object–this implies that 'plant' is a noun.

A written document is more than a set of words. The context in which each word occurs gives hints to the meaning of the word. To reduce the noise of multiple meanings, each word is tagged with its part of speech.

5.1 Brill's Tagger

Eric Brill introduced his rule-based part-of-speech tagger in 1992. Unlike stochastic taggers that require large manually annotated text to estimate lexical and contextual probabilities, Brill's tagger relies on a simple set of rules to assign part-of-speech tags. This method of tagging performs competitively with stochastic taggers, preserves learned data in human-readable rules instead of a matrix of statistics, and tags "ten times faster than the fastest stochastic tagger" [2, 2].

5.1.1 Transformation-Based Learning Algorithm

Brill's tagger is based on a transformation-based error-driven learning algorithm. The contextual and lexical rules included with Brill's tagger are ordered lists of rules. There are two components to Brill's rules, *rewrite rules* indicate a tag change and *triggering environments* indicate when to make a change [3, 545].

The learning algorithm applies these rules to annotated text to shift part-of-speech guesses closer to the truth. Brill's tagger uses these rules to implement the learning process demonstrated in Figure 5.1. The manually annotated corpus used to define the truth for the learner was the Penn Treebank Wall Street Journal Corpus.

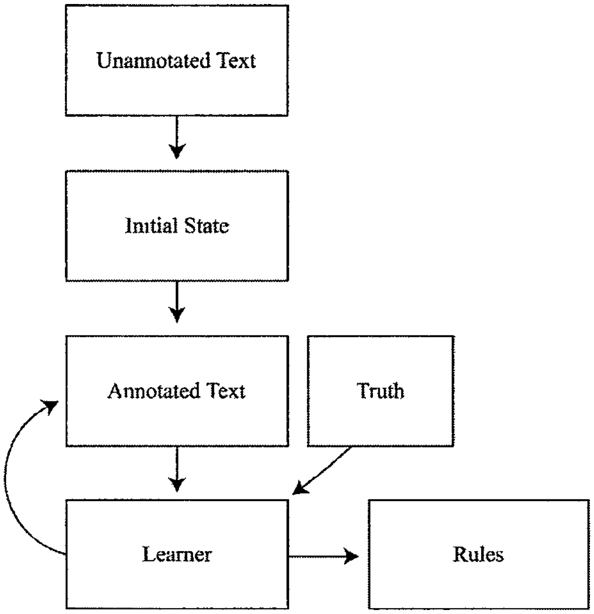


Figure 5.1 Transformation-Based Error-Driven Learning.

The same diagram that defines the transformation-based learning algorithm can be used to describe Brill's tagging process. *Unannotated text* refers to the original untagged text that is introduced into the system. The *initial state* assigns a value to the text. The text moves from the *initial state* to the *annotated text* state through the retrieval of the part-of-speech tag from the lexicon or the assignment of a default value if the text does not appear in the lexicon. The *annotated text* is compared with the *truth* in the *learning* state. The *truth* refers to the set of lexical and contextual rules that determine whether the tag should be changed given the syntactic properties of the text. Once all of the *annotated text* has been compared with all of the *truth*, *learning* stops. At this point, the tagging process is complete.

5.2 Components of Brill's Tagger

Brill's tagger consists of a lexicon that lists known words with their possible part-of-speech tags and two sets of rules that morphologically and contextually iteratively refine part-of-speech tag 'guesses'.

5.2.1 Lexicon

The lexicon included with Brill's tagger lists all possible part-of-speech tags for each word, with the most frequently occurring tag first, as seen in the Penn Treebank Wall Street Journal Corpus and the Brown Corpus. Figure 5.2 illustrates a few examples of the format of lexical entries in the Brill's lexicon.

```
wearing VBG
owe VBP VB
stimulators NNS
set VBN VBD VBP VBD|VBN JJ NN VB
midafternoon NN
Liqueur NNP
Casino NNP NN
foreclosed VBN JJ VBN|JJ VBD
disconnected VBN JJ
clipping NN
playoffs NNS
high-power JJ
Spook VBP
scourges NNS
double-jeopardy NN
Sheinberg NNP
plastic-covered JJ
decks NNS
frightened VBN JJ VBD
recitals NNS
shed VB VBD VBN VBP NN
relation-back JJ
```

Figure 5.2 Lexicon Entry Format.

5.2.2 Initial State

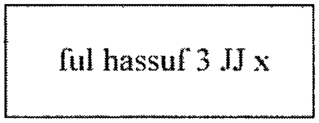
The tagging process begins by accepting at a sentence as unannotated text. Each word is looked up in the lexicon. If the word is found, the first tag in the word's part-of-speech list is assigned as the word's tag. If the word is not found, then the default tag for noun, NN (NNP if the word begins with a capital letter; NNS if the word is plural; or NNPS if it begins with a capital letter and is plural) is assigned to the word.

5.2.3 Lexical Rules

Once a sentence's preliminary tags are decided, lexical rules are applied to each word/tag pair in the annotated text to determine if a better tag exists for each word. The order of rule execution makes a difference. Transformations are directly influenced by every preceding transformation. Transformation rules indicate when to change tag *a* to tag *b*.

Lexical transformations are based on the word's morphological properties. The prefixes and suffixes of a word are evaluated to determine its best tag. If matches are found, tags are modified. This allows words that have either received a default noun tag or an inappropriate highest probability tag from the lexicon to be tagged correctly.

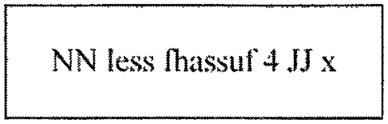
Figure 5.3 shows an example of a non-restricted lexical rule. Non-restricted rules change a words tag if the word meets the requirements–regardless of the word's current tag. This rule says to assign the tag 'JJ' to a word, if the last three letters are 'ful'.



```
ful hassuf 3 JJ x
```

Figure 5.3 Non-Restricted Lexical Rule.

Restricted lexical rules require a word's current tag to be a specific tag and the word to meet particular morphological requirements. Figure 5.4 demonstrates a restricted rule, that says that if the last four letters of a word are 'less' and the current tag is 'NN', then change the tag of the word to 'JJ'.



```
NN less fhassuf 4 JJ x
```

Figure 5.4 Restricted Lexical Rule.

5.2.4 Contextual Rules

Following lexical rule evaluation, a second set of rules is applied to the annotated text. These rules use surrounding words and their tags to adjust the tag of a given word. Contextual rules consider up to seven words at a time: three words to the left of a word, the word, and three words to right. This allows for the context of the word to play a role in determining its tag. Figure 5.5 shows an example of one of Brill's contextual rules.

This rule says that if the current tag is 'VBN', the previous tag is 'NN', and the next tag is 'DT' then change the current tag to 'VBD'.

```
VBN VBD SURROUNDTAG NN DT
```

Figure 5.5 Contextual Rule.

5.2.5 Penn Treebank Tagging Conventions

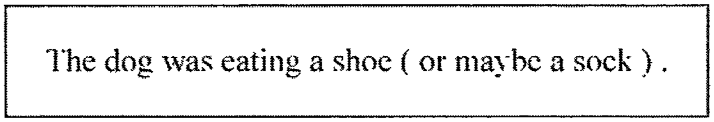
The tags assigned by Brill's tagger follow Penn Treebank tagging conventions. Words receive a two to four character tag. Penn Treebank tags are based on lexical rules and indicate specific qualities of words. Figure 5.6 shows a small subset of tags. For example, a word tagged JJR is not only an adjective, it is a comparative adjective that ends in '-er'. Many of Brill's lexical rules are directly derived from Penn Treebank's tagging conventions [7].

Part of Speech	Tag	Part of Speech (cont.)	Tag
Adjective	JJ	Modal verb	MD
Adjective, comparative	JJR	Numeral, cardinal	CD
Adjective, superlative	JJS	Possessive ending	POS
Adverb	RB	Proper noun, plural	NNPS
Article	DT	Proper noun, singular	NNP
Common noun, plural	NNS	Verb, base form	VB
Common noun, singular	NN	Verb, past tense	VBD
Conjunction, coordinating	CC	Verb, present participle	VBG

Figure 5.6 Frequently Occurring Parts of Speech with Associated Penn Treebank Tags.

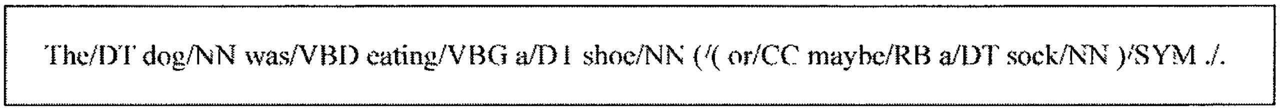
5.2.5 Tagging Example

Brill's tagger expects a sentence to be formatted with words and punctuation separated by spaces. This requires a pre-processing step. Figure 5.7 demonstrates the acceptable format for unannotated text. Figure 5.8 shows the resulting tagged sentence generated from the tagging process.



The dog was eating a shoe (or maybe a sock) .

Figure 5.7 Sentence Format for Input to Brill's Tagger.



The/DT dog/NN was/VBD eating/VBG a/DT shoe/NN (/(or/CC maybe/RB a/DT sock/NN)/SYM ./.

Figure 5.8 Tagged Sentence Returned from Brill's Tagger.

CHAPTER 6

DERIVING MEANING FROM CONTEXT

Once a word's part of speech has been identified, the set of potential meanings for a word is reducible to a subset of the word's total meanings. The next step is to identify what this word is *intended* to mean. This process should emulate the cognitive steps we take when we try to elicit meaning from context for unfamiliar words.

6.1 WordNet

WordNet was selected as a resource for defining the meaning of words in text. Beyond dictionaries and thesauruses, WordNet organizes words based on their meaning, not on their syntactic properties. The structure of WordNet's verb and noun conceptual hierarchies make it a compelling tool for deriving meaning via concept subsumption.

George Miller, founder of WordNet and the Program for Cognitive Studies at Princeton, has been working with the ideas of synsets (synonym sets) and mapping word forms to word meanings since 1985. Over the past twenty years, WordNet has evolved from a semantic net with synsets for forty-five nouns into a comprehensive network of 109,373 synsets for more than 107,000 nouns, 10,000 verbs and 25,000 modifiers [4].

Different parts of speech are categorized using different strategies. Nouns are represented in a hierarchy of 'is-a' relationships and describe the parts of an object; verbs are organized into clusters based on entailment relationships that define how an object functions; and adjectives and adverbs are stored as bipolar clusters.

6.2 Nouns and Meronymy

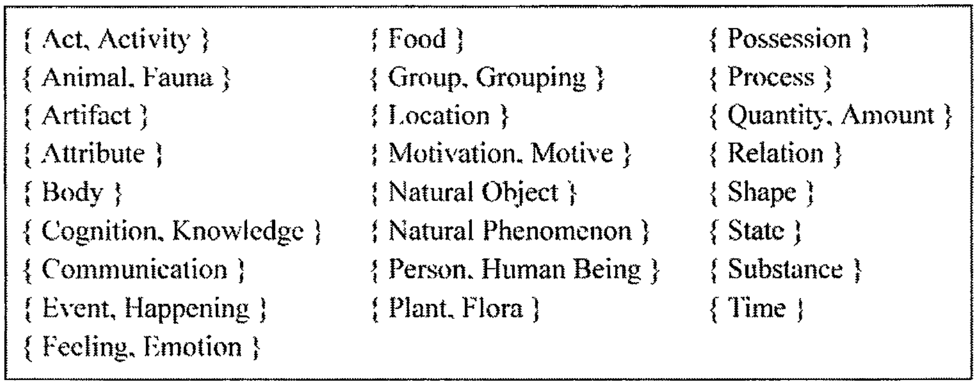
The hierarchical structure of nouns in WordNet relies heavily on meronym relationships. A *meronym* represents an 'is-part-of' relationship, such that *a* is a part of *b*. From the root of WordNet's hierarchy down to each leaf, the children of parent categories are synonyms that share a meronymous relationship with their parent. Similarly, from the bottom-up, all parent categories are *holonyms*–representing an 'is-composed-of' relationship with their children. There are three types of meronym relationships used by WordNet:

"W_m *#p* -> W_h indicates that W_m is a component part of W_h,

W_m *#m* -> W_h indicates that W_m is a member of W_h, and

W_m *#s* -> W_h indicates that W_m is the stuff that W_h is made from" [WNMIT,39].

Nouns are divided into several hierarchies, each having a unique beginner. WordNet version 1.5 identifies 25 unique beginner concepts representing distinct semantic categories, Figure 6.1, [4, 29]. Of these beginning concepts, further groupings were found which reduced the 25 beginners down to eleven. The top eleven categories are indicated in italics, in Figure 6.2 [4, 30].



{ Act, Activity }
{ Animal, Fauna }
{ Artifact }
{ Attribute }
{ Body }
{ Cognition, Knowledge }
{ Communication }
{ Event, Happening }
{ Feeling, Emotion }
{ Food }
{ Group, Grouping }
{ Location }
{ Motivation, Motive }
{ Natural Object }
{ Natural Phenomenon }
{ Person, Human Being }
{ Plant, Flora }
{ Possession }
{ Process }
{ Quantity, Amount }
{ Relation }
{ Shape }
{ State }
{ Substance }
{ Time }

Figure 6.1 WordNet 1.5, Beginning Noun Concepts.

To determine the current noun hierarchy in WordNet 1.6, an actual enumeration of the top 5 levels of WordNet's noun hierarchy was created (see Appendix A). This was achieved by recording successive hyponyms for every concept starting with the root concept, 'Entity'. Successive hyponym calls revealed that WordNet 1.6 uses a different set of noun beginners than WordNet 1.5, see Figure 6.3. In fact, all of the noun concepts are now drawn under two unique beginners: Physical Entity and Abstract Entity. The third category "Thing" contains a flat collection of eight seemingly unrelated terms.

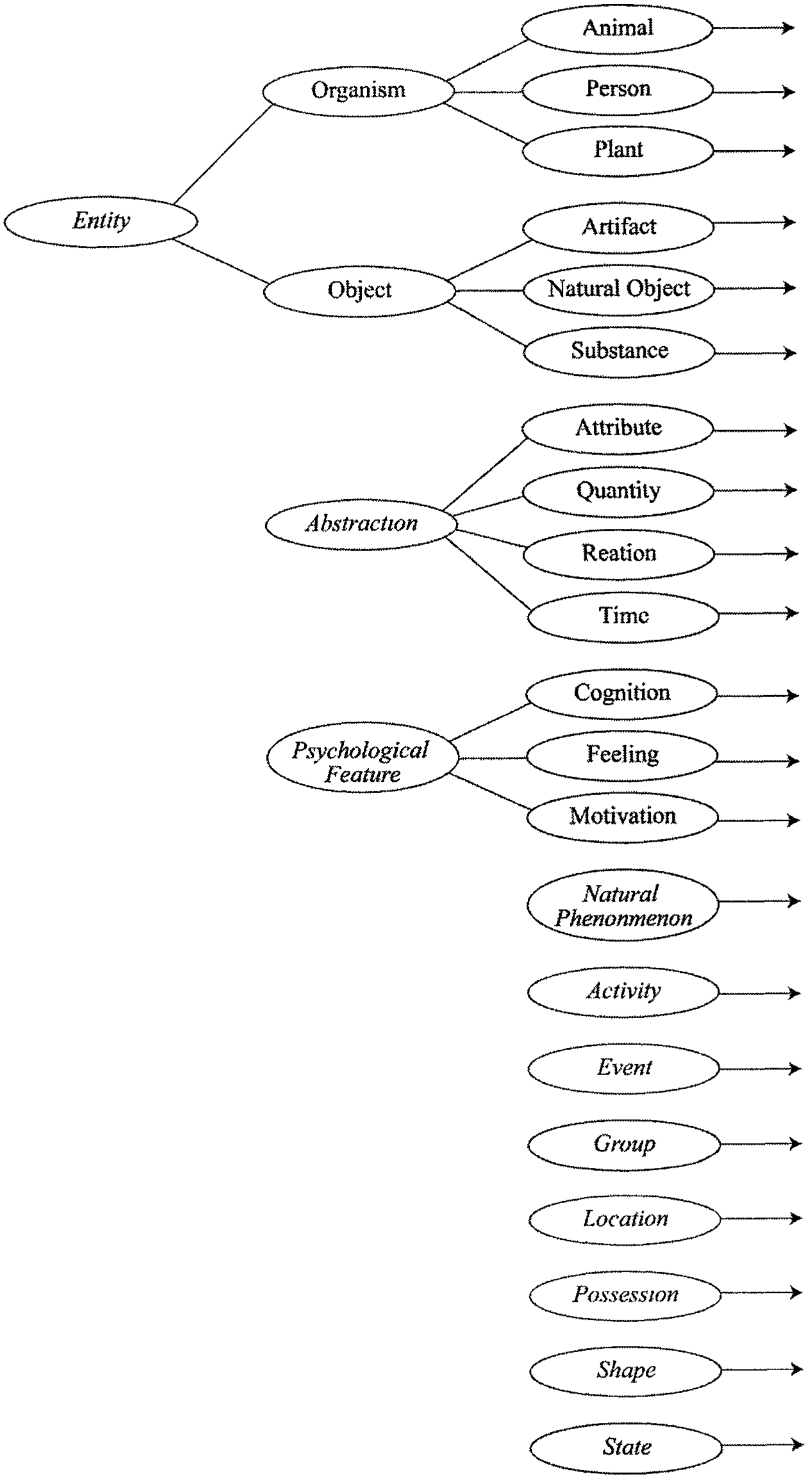


Figure 6.2 WordNet 1.5, Reduced Set of Beginning Noun Concepts.

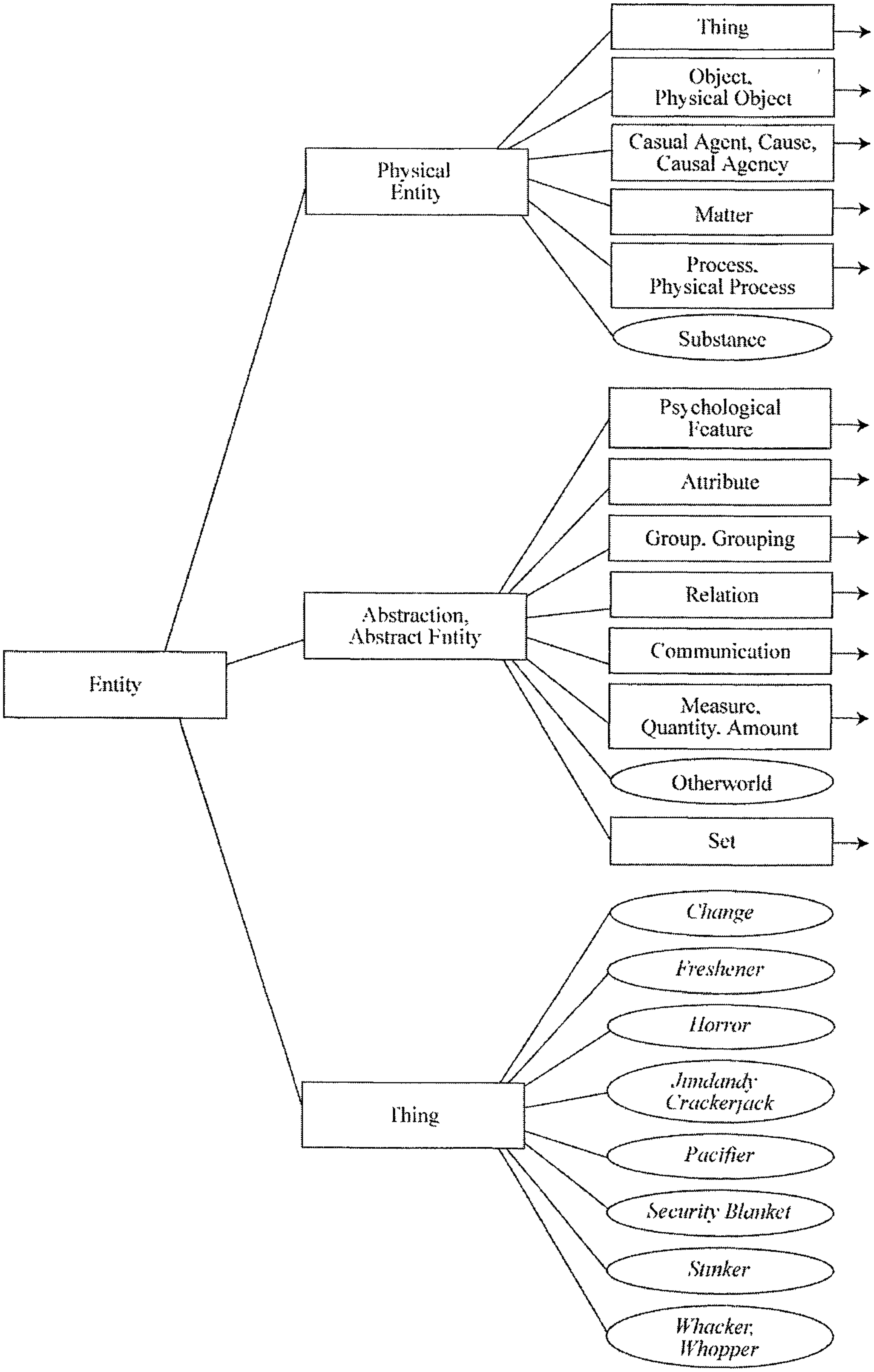


Figure 6.3 WordNet 1.6, Revised Set of Beginning Noun Concepts.

6.3 Verbs and Entailment

Verbs are categorized similarly to nouns in a verb hierarchy of synsets. The distinguishing aspect of verb organization ıs the use of the *entailment* relatıon. In simplest terms, the entailment relation between the two verbs V_1 and V_2 holds if the truth of V_1 requires the truth of V_2. For example, eat entails hunger–as eating requires hunger. Negation reverses the direction of entailment, i.e. not eating entails not hungry [4, 77]. There are four types of entailment relationships between verbs in WordNet, see Figure 6.4 [4, 84].

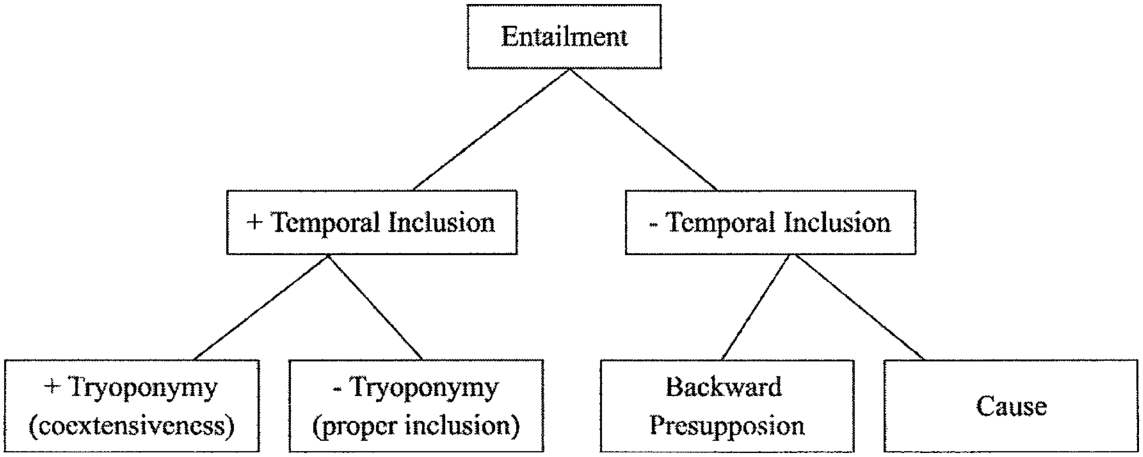
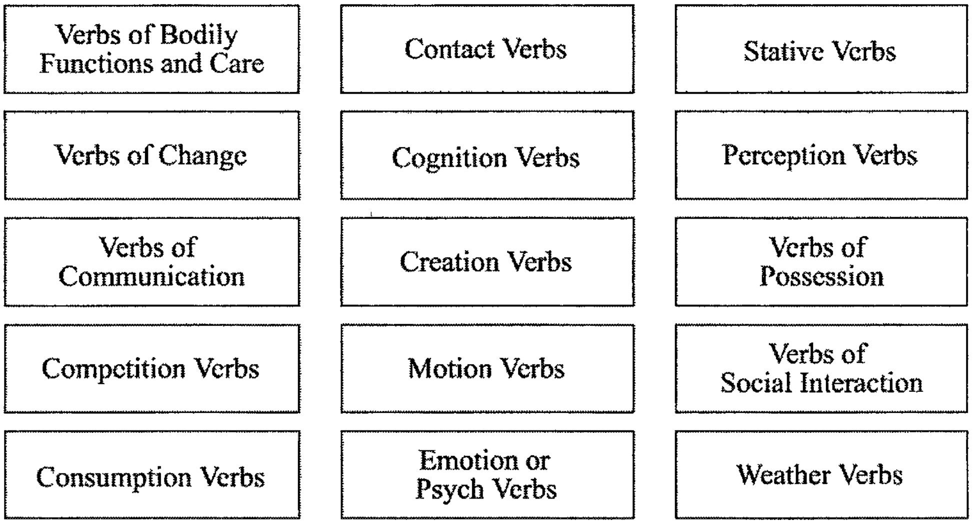


Figure 6.4 Entailment Relationships of Verbs in WordNet.

Verbs are dıvided into fifteen types of categories, shown in Figure 6.5. Important to note–the edges between categories blur due to polysemy and not all verbs fit nicely in this structure. Also noteworthy, the combınation of fifteen unique concept category beginners and the inability of verbal relations to form part-whole bonds leads to shorter paths to verbs in WordNet than nouns.

Stative verbs that are various forms of the concept "be" (*is*, *resemble*, *enough)*; *control* verbs (*need*, *must*, *succeed)*; and *aspectual* verbs (*start, begin)* are stored in an auxiliary class outside of the semantic verb hierarchy [9, 57-61].



Verbs of Bodily Functions and Care	Contact Verbs	Stative Verbs
Verbs of Change	Cognition Verbs	Perception Verbs
Verbs of Communication	Creation Verbs	Verbs of Possession
Competition Verbs	Motion Verbs	Verbs of Social Interaction
Consumption Verbs	Emotion or Psych Verbs	Weather Verbs

Figure 6.5 Verb Categories ın WordNet.

Decomposing actions into 'sub-actions' is a different process than decomposing objects into parts, as 'sub-actions' aren't always distinguishable units. The entailment relation ıs used to break verbs down into their smaller parts using *time* as a distinguishing characteristic. An action is seen as a sequence of events occurring over time. There are two types of entailment: those that require temporal inclusion and those that don't.

One type of entailment relation that requires temporal inclusion is *troponymy*– which refers to a method of doıng something. "Every troponym V_1 of a more general verb V_2 also entails V_2"[9, 47]. Troponyms are *temporally coextensive*, meaning that V_1 must occur at the same time as V_2. Whispering is a *troponym* of speaking. He is whispering *entails* he is talking, and *every moment* that he is whispering, he is talking.

The second type of entailment relation that requires temporal inclusion is *proper inclusion*. "A verb V_1 will be said to include a verb V_2 if there is some stretch of time during which the activities denoted by the two verbs co-occur, but no time during which V_2 occurs and V_1 does not" [4, 78]. Dance *properly includes* spin and sleep *properly includes* snore.

There are two types of entailment relations for verbs that aren't bound by temporal inclusion. *Backward presupposition* is a form of entailment that relates two verbs in a 'result-of' or 'purpose' relation. "A verb V_1 that is entailed by another verb V_2 via backward presupposition cannot be said to be a part of V_2" [9, 51]. Entailment relations do not share a part-whole bond. 'Pass' *entails* 'test' by *backward presupposition.*

The *cause* relation is the last type of entailment relation recognized by WordNet. "If V_1 necessarily causes V_2, then V_1 also entails V_2.[9, 54]. This relation is complicated and only holds in WordNet for cause/result pairs when "the synonyms of the members of such a pair inherit the Cause relation, indicating that this relation holds between the entire concept rather than between individual word forms only: the synonyms {teach, instruct, educate}, for example, are all causatives of the concept {learn, acquire knowledge}" [9, 53].

6.4 Modifiers and Bipolar Clusters

Modifiers include adverbs and adjectives. The function of modifiers is to decribe nouns. WordNet defines adjectives into two categories: descriptive adjectives and relational adjectives. Descriptive adjectives are by far the largest category of modifiers. They function as typical adjectives, in that they simply describe nouns–i.e. *large*, *stinky*, and *intelligent.*

Descriptive adjectives are organized using the antonymy relation between clusters of synsets. Figure 6.6 illustrates how sets of synonyms are tied to their complementary set of synonyms, in bipolar clusters. “Most antonyms of descriptive adjectives are formed by a morphological rule that changes the polarity of the meaning by adding a negative prefix” [4, 49]. Binary antonym pairs like up/down, left/right, living/dead as well as gradable antonyms like bright, light, dim, dark fit naturally into this type of organization. There are more than 1700 clusters for antonym pairs in WordNet.

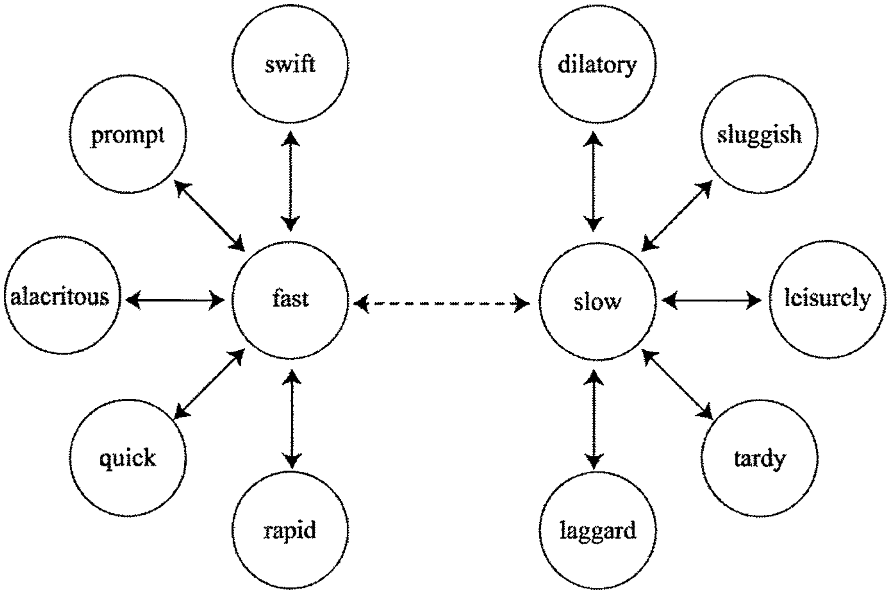


Figure 6.6 Bipolar Adjective Structures in WordNet.

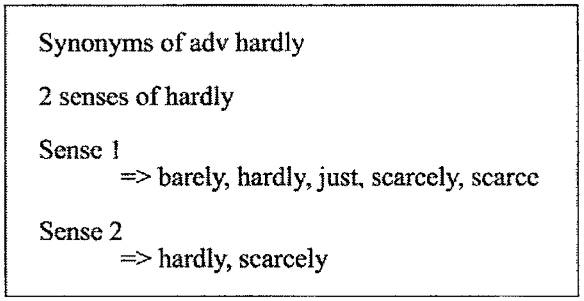
To accommodate different meanings for the same word form, word forms are numbered. For example ‘light’ meaning ‘bright’ will be labeled distinctly from ‘light’ meaning ‘not heavy’. If the former was labeled *light1,* then the *latter* would be labeled with a different number, for example, *light2.*

When antonyms can’t be found for words, *similarity pointers* are used to link a word with no antonym to a similar word with an antonym. Similarity pointers represent indirect antonyms through an ‘is-similar-to’ relation. An exception to this strategy is colorful words that represent extreme concepts. Words like *angry* or *excited.* Nearly similar words miss the passion of the concept, displeased hardly equates angry. These kinds of extreme words are paired with the “non” version of themselves, creating a “non” antonym pair, i.e. angry is matched with not angry.

Relational adjectives act like modifying nouns and function as classifiers. [WordNet, 59]. Theses kinds of adjectives don’t function like descriptive adjectives and don’t fit well in the bipolar cluster arrangement. Because of their functional resemblance to nouns, relational adjectives that don’t have a logical antonym are stored separately in a relational adjective file. The adjectives in this file are linked to their corresponding noun synsets. There are 2,832 relational adjective synsets in WordNet 1.5.

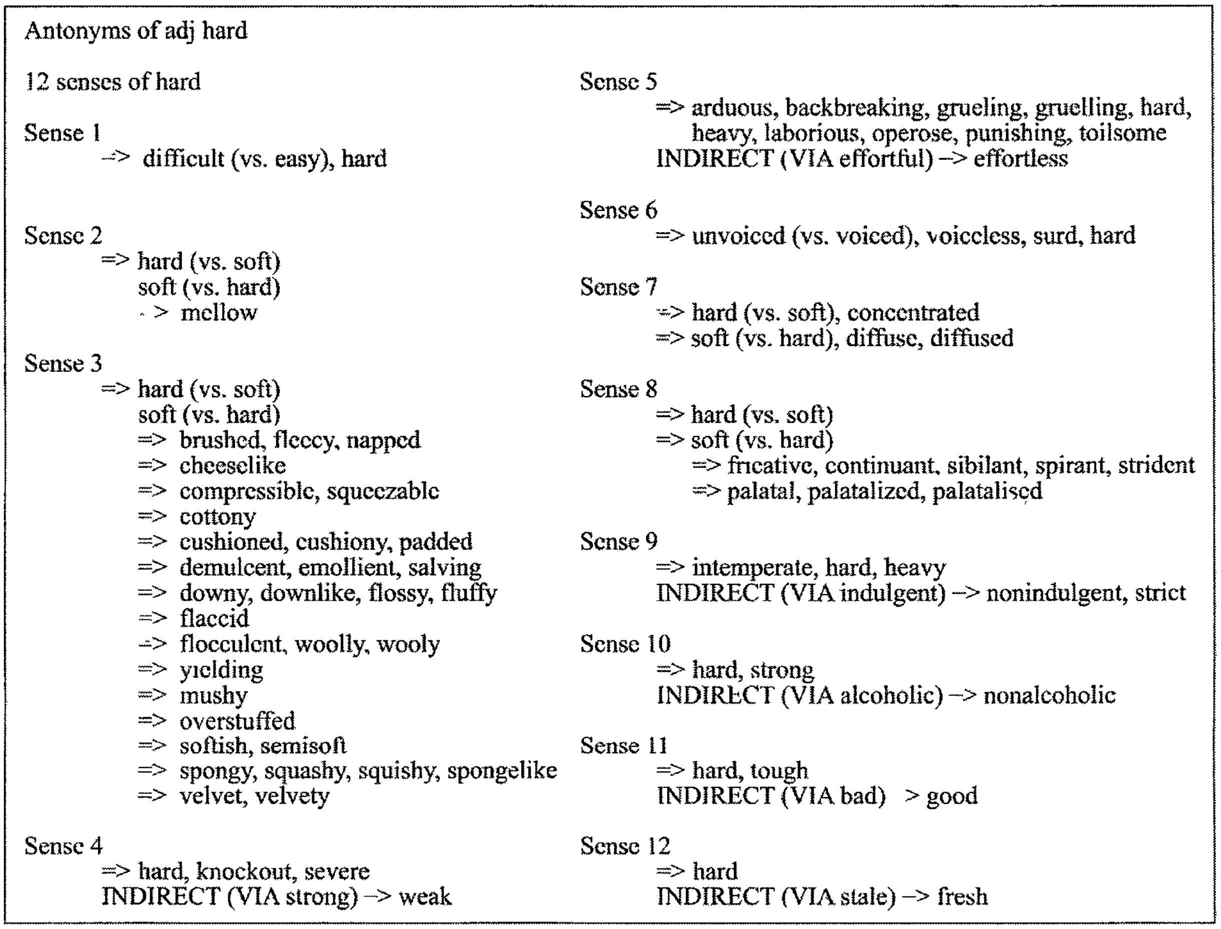
Adverbs are linked to adjectives with a ‘derived-from’ relation. There is a tight correlation between many adverbs and their base adjective, for example *slow* and *slowly.* Adverbs like this are directly linked to their adjective counterparts. However this isn’t a sweeping strategy, as some adverbs have nothing to do with their adjective base, consider *hardly* and *hard.*

Because adverbs behave so erratically, the organization of adverbs in WordNet is basically flat. Only appropriate synonym and antonym relations are accommodated. Figure 6.7 shows the synset for *hardly,* because it unrelated to the adjective *hard,* there is no referential pointer to its base adjective. Figure 6.8 shows the synset for *hard.*



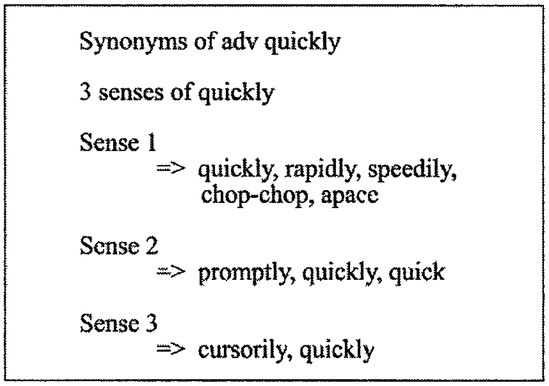
```
Synonyms of adv hardly

2 senses of hardly

Sense 1
        => barely, hardly, just, scarcely, scarce

Sense 2
        => hardly, scarcely
```

Figure 6.7 Example of Adverb Synset for 'hardly'.



```
Antonyms of adj hard

12 senses of hard

Sense 1
      => difficult (vs. easy), hard


Sense 2
      => hard (vs. soft)
         soft (vs. hard)
         => mellow
Sense 3
      => hard (vs. soft)
         soft (vs. hard)
         => brushed, fleecy, napped
         => cheeselike
         => compressible, squeezable
         => cottony
         => cushioned, cushiony, padded
         => demulcent, emollient, salving
         => downy, downlike, flossy, fluffy
         => flaccid
         => flocculent, woolly, wooly
         => yielding
         => mushy
         => overstuffed
         => softish, semisoft
         => spongy, squashy, squishy, spongelike
         => velvet, velvety

Sense 4
      => hard, knockout, severe
      INDIRECT (VIA strong) -> weak

Sense 5
      => arduous, backbreaking, grueling, gruelling, hard,
         heavy, laborious, operose, punishing, toilsome
      INDIRECT (VIA effortful) -> effortless

Sense 6
      => unvoiced (vs. voiced), voiceless, surd, hard

Sense 7
      => hard (vs. soft), concentrated
      => soft (vs. hard), diffuse, diffused

Sense 8
      => hard (vs. soft)
      => soft (vs. hard)
         => fricative, continuant, sibilant, spirant, strident
         => palatal, palatalized, palatalised

Sense 9
      => intemperate, hard, heavy
      INDIRECT (VIA indulgent) -> nonindulgent, strict

Sense 10
      => hard, strong
      INDIRECT (VIA alcoholic) -> nonalcoholic

Sense 11
      => hard, tough
      INDIRECT (VIA bad) -> good

Sense 12
      => hard
      INDIRECT (VIA stale) -> fresh
```

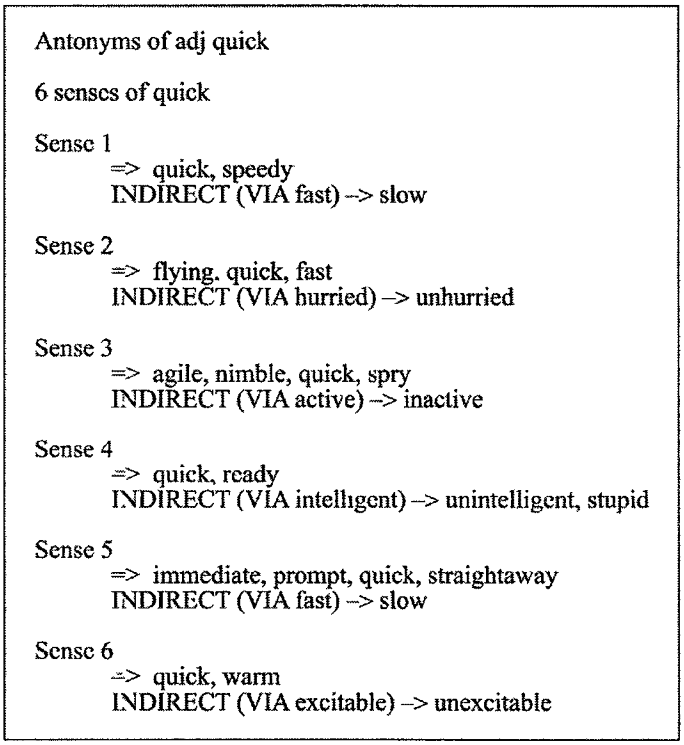
Figure 6.8 Example of Adjective Synset for 'hard'.



```
Synonyms of adv quickly

3 senses of quickly

Sense 1
        => quickly, rapidly, speedily,
           chop-chop, apace

Sense 2
        => promptly, quickly, quick

Sense 3
        => cursorily, quickly
```

Figure 6.9 Example of Adverb Synset for 'quickly'.

In a second example, Figure 6.9 shows the synset for *quickly*. In this case, the second synset, 'Sense 2' contains a reference to its adjective base, *quick*. Figure 6.10 shows the synset for the adjective quick.



```
Antonyms of adj quick

6 senses of quick

Sense 1
        => quick, speedy
        INDIRECT (VIA fast) --> slow

Sense 2
        => flying. quick, fast
        INDIRECT (VIA hurried) --> unhurried

Sense 3
        => agile, nimble, quick, spry
        INDIRECT (VIA active) --> inactive

Sense 4
        => quick, ready
        INDIRECT (VIA intelligent) --> unintelligent, stupid

Sense 5
        => immediate, prompt, quick, straightaway
        INDIRECT (VIA fast) --> slow

Sense 6
        => quick, warm
        INDIRECT (VIA excitable) --> unexcitable
```

Figure 6.10 Example of Adjective Synset for 'quick'.

CHAPTER 7

PROCESS

The process used in identifying similarity within a set of documents has been modeled to simulate the way humans determine similarity. Feature identification is a human method of understanding an object or an action. To this end, documents are processed to become collections of features. The human mind seeks to categorize new information based on its general understanding of the world. To this end, words are replaced with semantic pathways that simulate understanding. Understanding is expressed as a conceptual framework. The human brain relies on the comparison of features for identifying and integrating new information and assessing similarity. Similarity between two documents becomes a comparison of their common and distinctive features realized by comparing the conceptual framework of one document to another.

7.1 Leveling the playing field

The initial task ın document similarity assessment is to translate a set of documents into a set of comparable feature sets. In this case, the documents have been gathered from the Reuters-21578 document set, described in chapter three. The corpus of documents used in this process has been randomly selected from the Reuters-21578 document set. These files are organized in five data sets–each with fifty documents.

The number of features defined per document is directly related to the word count of a document. There is wide variation in the number of words per Reuters-21578 document, some have as few as three words while others have thousands of words. This variation is why there must be at least 50 words per document.

7.2 Resolving Ambiguity

There is a twofold approach to resolving ambiguity. The initial step reduces the scope of possible meanings that a word may have. This is done by only considering the meaning of a word given its part of speech. If multiple meanings are found given a word's particular part of speech tag, the search space is maximized and all possible meanings are considered. This greedy strategy increases the opportunity for the correct meaning to emerge when weighted by context.

7.2.1 Lexical Ambiguity

Polysemous words are those with multiple meanings. Lexical analysis considers part-of-speech tags to reduce the total number of meanings a word may have. Consider the fragments: 'I'm going to plant cotton this year'; and 'I work at a cotton plant'. In the first instance, plant is a verb and cotton is a noun–in the second, cotton is an adjective and plant is a noun. Many cases like these can be clarified through part-of-speech verification. Therefore, part-of-speech tags are used to reduce the number of meanings a word may have.

Second, multiple senses for each word are preserved. When words can mean multiple things–given a particular part of speech, the context is allowed to weight the meaning. If a word has eleven meanings–it has eleven word senses and eleven edge lists.

Each path from word sense to root is discovered by queries to WordNet. These paths are included in the EdgeNode list and the semantic network of a document. This allows all meanings to be considered, increasing the odds for similarities to emerge.

7.2.2 Semantic Ambiguity

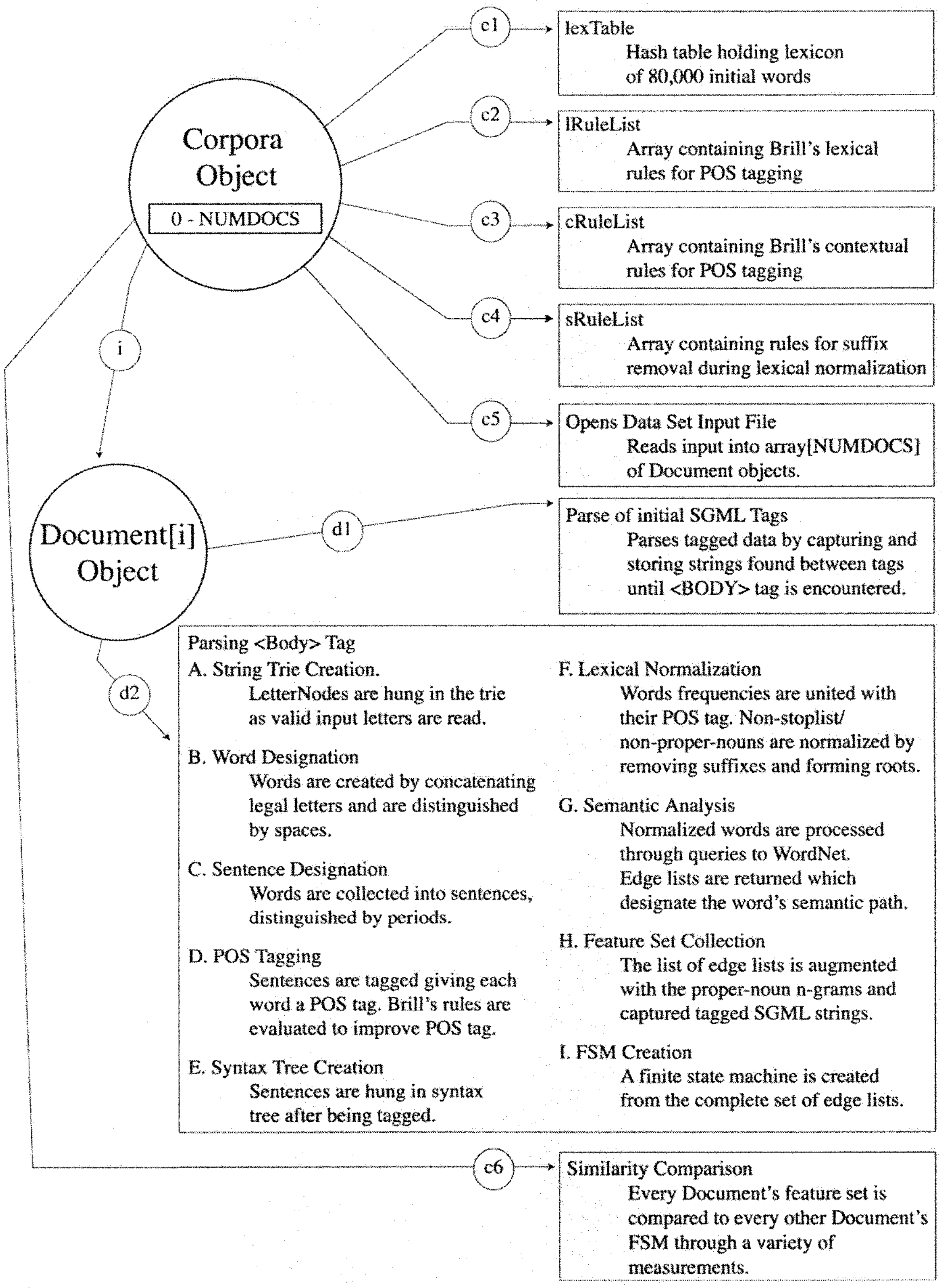
Many word sequences contain ambiguous words that refer to more than one concept. This type of ambiguity is fodder for jokes like "Is your refrigerator running?"– where running could mean 'operating' or 'moving quickly'. This type of semantic ambiguity will be largely resolved by the inclusion of all possible meanings for the word "running", given the most likely part-of-speech tag 'VB'. More drastic examples, where the words themselves no longer reflect the intended meaning, like the phrase "burning the midnight oil," are beyond the scope of this thesis.

7.2.3 Syntactic Ambiguity

Some word sequences can be interpreted in more than one way. The statement "I need a hand." can be interpreted in a variety of ways. The most likely meaning is that someone requires assistance. But, if the neighboring context was thespian related, it could mean that someone would like applause–or if the context was a surgical setting, someone could be doing an operation and require an anatomical human hand.

Rather than choose a single sense of a word, which may be wrong, all senses are considered. When comparisons are made between documents, overlapping semantic paths resolve ambiguity. This follows WordNet precedence. By providing all possible word senses in the semantic network of a document, context is expanded to cover all possible meanings.

7.2 Overview



lexTable
Hash table holding lexicon of 80,000 initial words

lRuleList
Array containing Brill's lexical rules for POS tagging

cRuleList
Array containing Brill's contextual rules for POS tagging

sRuleList
Array containing rules for suffix removal during lexical normalization

Opens Data Set Input File
Reads input into array[NUMDOCS] of Document objects.

Parse of initial SGML Tags
Parses tagged data by capturing and storing strings found between tags until <BODY> tag is encountered.

Parsing <Body> Tag

A. String Trie Creation.
LetterNodes are hung in the trie as valid input letters are read.

B. Word Designation
Words are created by concatenating legal letters and are distinguished by spaces.

C. Sentence Designation
Words are collected into sentences, distinguished by periods.

D. POS Tagging
Sentences are tagged giving each word a POS tag. Brill's rules are evaluated to improve POS tag.

E. Syntax Tree Creation
Sentences are hung in syntax tree after being tagged.

F. Lexical Normalization
Words frequencies are united with their POS tag. Non-stoplist/ non-proper-nouns are normalized by removing suffixes and forming roots.

G. Semantic Analysis
Normalized words are processed through queries to WordNet. Edge lists are returned which designate the word's semantic path.

H. Feature Set Collection
The list of edge lists is augmented with the proper-noun n-grams and captured tagged SGML strings.

I. FSM Creation
A finite state machine is created from the complete set of edge lists.

Similarity Comparison
Every Document's feature set is compared to every other Document's FSM through a variety of measurements.

Figure 7.1 Overview of Process.

Figure 7.1 shows an overview of the process developed to assess similarity. These are the top-most phases and are labeled *c1* through *c6*. Phase *c5* includes the creation of every Document object. The phases *d1*, *d2* and *A-I* give an overview of the second tier of phases. Second tier phases are directly related to processing Document objects.

Given the SGML format of these documents, the values of the tagged attributes TOPIC, PLACES, PEOPLE, ORGS, EXCHANGES, COMPANIES, TITLE, and BODY are naturally included as members of the feature set. Each of these features contains zero or more attribute values, depending on how many values are present in the document. While the tagged attributes are not used in semantic analysis, they are considered members of the document feature set and are included in similarity comparisons.

The text encapsulated in the BODY tag ıs decomposed and evaluated to determine: words, word frequencies; part-of-speech tags; and proper noun n-gram sequences. Semantic features, derived from non-proper/non-stoplist nouns and verbs take the form of edge lists relating their hierarchical position in WordNet.

A single word may have multiple senses in WordNet. Each word sense returns a unique edge list. The inclusion of multiple senses increases the size of the semantic network and allows for a wider possible variety of concept subsumption. In other words, it casts a wider net.

Once all of the edge lists are identified, the features of the document are combined into two structures: an EdgeNode list and a finite state machine. The FSM functions as the semantic network and the edge-list becomes the complete feature set for the document.

The EdgeNode list originally contains only the edge lists returned from WordNet queries. This list is transformed into the feature set of the document with the addition of the proper noun n-grams and captured data from the SGML tags. Once the EdgeNode list contains a complete feature set, the FSM is created.

Each element in the edge-list is decomposed into a sequence of states–ending edges are designated as accepting states. The states are linked together to create the finite state machine. Similarity becomes a measure of common versus distinct states derived from comparing the edge-list of one document to the FSM of another.

7.3 Parsing SGML

SGML tagged data for the tags: TOPICS, PLACES, PEOPLE, ORGS, EXCHANGES, COMPANIES, TITLE and DATELINE is captured as-is. These tags are stored in Document objects in their associated field. For example, strings read from the topics tag are stored in the Document string topicsTag. Strings associated with the other tags occurring before BODY are ignored because the data is irrelevant.

Once the BODY tag is read, the parsing becomes much more sensitive. The set of recognized characters is reduced. Only these characters are accepted by Corpora's parse: A-Z, a-z, 0-9, the comma, the apostrophe, the period, the hypen and the quotation mark. Of these, characters only A-Z, 0-9. the apostrophe, the hyphen, the period, and the space are used in constructing the string trie. The recognized characters within the BODY tag become the letters in the string trie and the recognized words in the document.

The space character is used to designate the ending of words. It is the only character that LetterNode objects accept but never receive spaces directly from input. The only time a LetterNode receives a space is when it is inserted specifically to separate proper nouns in proper-noun n-gram sequences.

The appearance of the word REUTERS or Reuters indicates that the end of the BODY tag is coming up next. With the read of the closing tags /BODY, /TEXT, and /REUTERS, Corpora moves to the next Document object in the Document array and begins the steps of SGML processing again.

Once the end of file tag is read, or the maximum number of documents has been reached, similarity comparison between all of the documents begins.

7.4 Part-of-Speech Tagging

Tagging is performed primarily to reduce ambiguity. The tagging phase of this process has been largely based on Eric Brill's rule-based part-of-speech tagger, described in chapter five.

The code implementing Brill's rule sets and lexicon has been completely rewritten. The original code recreated the lexicon and rule sets on every call, for each of the three phases of tagging. For a data set of 10 documents, each with ten sentences, the lexicon and rule sets would be recreated three hundred times. The lexicon and rule sets are now read and stored one time per data set.

Brill's lexical rule set has been slightly modified. Rules explicitly labeling integers, conjunctions, modal verbs, the special case 'to', and determiners have been added. The rules are shown in Figure 7.2. They were added because there was a small number of words that exactly described each category.

```
9 char CD x
8 char CD x
7 char CD x
6 char CD x
5 char CD x
4 char CD x
3 char CD x
2 char CD x
1 char CD x
0 char CD x

and isconjun CC x
but isconjun CC x
or isconjun CC x
nor isconjun CC x
yet isconjun CC x
plus isconjun CC x
minus isconjun CC x
less isconjun CC x
times isconjun CC x
because isconjun CC x

to specialtag TO x

can ismodal MD x
could ismodal MD x
dare ismodal MD x
may ismodal MD x
might ismodal MD x
ought ismodal MD x
shall ismodal MD x
should ismodal MD x
will ismodal MD x
would ismodal MD x

all isdet DT x
both isdet DT x

the isdet DT x
every isdet DT x
no isdet DT x
a isdet DT x
an isdet DT x
either isdet DT x
neither isdet DT x
that isdet DT x
these isdet DT x
this isdet DT x
those isdet DT x
some isdet DT x
each isdet DT x
```

Figure 7.2 Added Lexical Rules.

The process of tagging closely follows the process of Brill's tagger. Only slight modifications have been made, and these are the result of differing coding styles and implementation environment.

As sentence endings are determined, sentences are translated from strings into SyntaxNode sub-trees. The three phases of tagging are implemented by looking up each word in the lexicon and assigning the first tag value; lexically refining the tags of each word by applying each of the rules in the lexical rule set to each word in the sentence; and then further refined by examining the context of the word's tag compared to the surrounding tags.

After a sentence is tagged, each word with its part-of-speech tag is transferred to a new SyntaxTree node. The sub-tree sentence root's array preserves the original order of the words in each sentence. Once the sub-tree is complete, the sentence is hung in the appropnate position of its parent paragraph SyntaxTree node.

7.5 Lexical Normalization

Lexical Normalization is the process of reducing the number of distinct words that represent a document. After a document is read, its string trie is examined in a depth-first traversal. As each word is reached in the traversal, its Location list is searched. Each Location node contains a set of coordinates into the syntax tree.

A	ABOUT	AFTER	ALONG
ALSO	ALTHOUGH	AN	AND
ANOTHER	ANY	ARE	AROUND
AS	AT	BE	BEEN
BECAUSE	BOTH	BUT	BY
CAN	COULD	DID	DO
DOES	DOWN	FOR	FROM
FURTHER	GONE	HAD	HAS
HAVE	HAVING	HE	HERS
HIS	HOW	HOWEVER	IF
IN	INTO	IS	IT
ITS	MORE	MOST	MUCH
NO	NOT	NOW	OF
OFF	ON	ONE	OR
OTHER	OTHERS	OVER	PER
REUTERS	SEEN	SAID	SHE
SHOULD	SOME	STILL	THAT
THE	THEIR	THERE	THESE
THEY	THIS	THOSE	THROUGH
TO	TOO	UNDER	UP
WERE	WAS	WAY	WE
WHAT	WHEN	WHERE	WHICH
WHILE	WHY	WILL	WITH
WOULD	YET		

Figure 7.3 Stoplist Words.

Lexical coordinates are used as parameters to retrieve the part-of-speech tag assigned to the word in a given syntax tree position. The tags of each word are concatenated into a collective tag list. The frequency of the word is compared with the elements in the tag list to verify correctness.

All words that are not proper nouns and are not members of the stoplist word group are selected for lexical normalization. Stoplist words are words that are removed from consideration during processing because they occur frequently. Words common to all documents lack diagnostic value. Words considered stoplist words are shown in Figure 7.3.

Each word and its list of locations is passed into a function that creates the set of words that are eventually considered during semantic analysis. This function takes a word with its list of tags, Figure 7.4, removes suffixes and divides the word based on its tags, Figure 7.5.

Word: *word1*   Tags: *t1*, *t2*, *t1*, *t1*

Figure 7.4 Word with List of POS Tags.

Word: *word1*   Tags: *t1*   Frequency: *3*
Word: *word1*   Tags: *t2*   Frequency: *1*

Figure 7.5 Word and Frequency Based on POS Tag.

The result is a list that contains the number of times each root word in the document occurs given a particular part-of-speech tag.

7.5.1 Suffix Removal

This use of rules for suffix removal was greatly influenced by the efficacy of Brill's rule sets. Figure 7.6 demonstrates the rules created for suffix removal. Based on the grammatical rules learned in elementary school, these rules were created to cover the majority of cases. These rules are extendable for future development.

Special cases, words that don't follow the rules, have exceptions created for them in the code. These are words like "cookies" and "patrolled". Not all special cases can be accounted for, but an attempt to cover frequently occurring cases has been made.

Removing suffixes allows for preliminary semantic analysis–in that word forms and their tags can be mined for intrinsic meaning. The word 'leaves' with a 'verb' part-of-speech tag distinguishes it from 'leaves' tagged as a 'noun'–one reduces to the root 'leaf' and the other reduces to 'leave'. This initial data mining can narrow down the scope of a words meaning and further efforts to reduce ambiguity.

7.5.2 Frequency Counts

The nodes in a document's string trie are used to count word frequencies during the original read of a document. A single post-order traversal of the string trie is sufficient to retrieve words for lexical normalization by generating an alphabetically ordered list of unique elements complete with their associated frequencies.

3 IED Y	{ last 3 letters are 'ied', remove 'ied' add 'y'} *ex. readied -> ready*
3 IES Y	{ last 3 letters are 'ies', remove 'ies' add 'y'} *ex. puppies -> puppy*
3 IER Y	{ last 3 letters are 'ier', remove 'ier' add 'y'} *ex: cheerier -> cheery*
3 ING 2 VC +E	{ 1 vowel then 1 consonant before 'ing', delete 'ing' and 'e' } *ex: leaving -> leave*
3 ING 2 CC -C	{ 2 same consonants before 'ing', delete 'ing' and the last consonant} *ex hopping -> hop*
3 ING 2 CD	{ 2 different consonants before 'ing', delete 'ing' } *ex. resting -> rest*
3 SEN SE	{ last 3 letters are 'sen', remove 'n'} *ex. loosen -> loose*
3 SER SE	{ last 3 letters are 'ser', remove 'r'} *ex. cleanser -> cleanse*
3 VES F NN	{ last 3 letters are 'ves', part-of-speech tag is NN, remove 'ves' add 'f' } *ex: shelves -> shelf*
3 VES VE VB	{ last 3 letters are 'ves', part-of-speech tag is VB, remove 's' } *ex leaves -> leave*
3 VEN VE	{ last 3 letters are 'ven', remove 'n' } *ex: driven -> drive*
2 ER 3 CCC	{ 3 consonants before 'er', delete 'er' } *ex: lighter -> light*
2 EN 2 CC -C +E	{ 2 same consonants before 'en', delete 'en' and last consonant, add E } *ex hidden -> hide*
2 ER 2 CC -C	{ 2 same consonants before 'er', delete 'er' and last consonant} *ex: hopper -> hop*
2 EN 3 CCC	{ 3 same consonants before 'en', delete 'en' } *ex: lighten -> light*
2 EN 3 VVC	{ 2 vowels then 1 consonant before 'en', delete 'en' } *ex: claimen -> claim*
2 EN 3 VLL -EN	{ 1 vowel then 2 "L"s before 'en', delete 'en' } *ex: fallen -> fall*
2 EN 3 VCC -CEN -E	{ 1 vowel then 2 same consonants before 'en', delete last consonant and 'en', add 'e' } *written -> write*
2 EN 3 CVC -N	{ 1 consonant then 1 vowel then 1 consonant before 'en', delete 'n' } *ex. biten -> bite*
2 ED 3 VCC -C	{ 1 vowel then 2 same consonants before 'en', delete last consonant and 'ed' } *ex incurred -> incur*
2 ED 3 VVC	{ 2 vowels then 1 consonant before 'ed', delete 'ed' } *ex: aimed -> aim*
2 ED 3 CVC +E	{ 1 consonant then 1 vowel then 1 consonant before 'ed', delete 'd' } *ex: fined -> fine*
2 ER	{ default 'er', delete 'er' } *ex: clearer -> clear*
2 ED	{ default 'ed', delete 'ed' } *ex: famished -> famish*
1 S	{ default 's', delete 's' } *ex. plays -> play*

Figure 7.6 Suffix Removal *Rules.*

At the end of Lexical Analysis, the initial list of words and their frequencies are transformed into a list of non-stoplist, non-proper-noun, root words. Each root word is grouped with its associated part-of-speech tag and frequency. This revised list becomes the key structure for determining the semantic meaning of the document. The frequencies eventually become modifiers of concept strength used in similarity comparison.

7.6 Semantic Analysis

At this point in the process, a document has been reduced to its effective feature set. This set consists of three distinct types of structures: a list of words prepared for semantic analysis, a set of extracted tagged topic data, and a list of proper noun n-grams. With the features identified, the focus shifts to evaluating semantics.

The hierarchies of nouns and verbs in WordNet are the most compelling for similarity comparison. These structures are well suited for creating conceptual pathways. Modifiers, however are stored in bipolar clusters. There is no conceptual hierarchy available for modifiers and little is gained beyond flat synsets of antonyms and synonyms. For this reason, during semantic analysis the list of words is again reduced. From this point on, only nouns and verbs are considered.

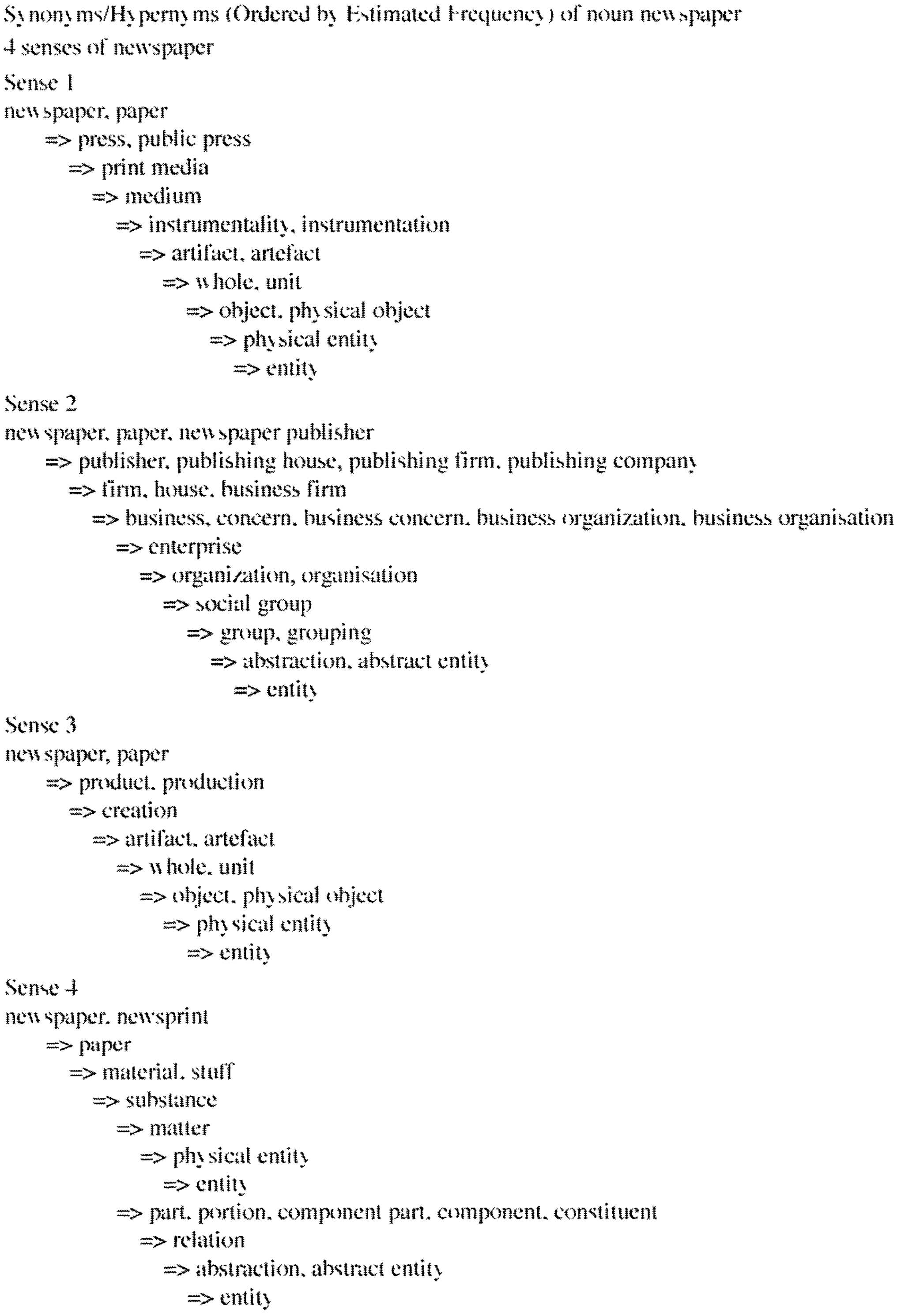
7.6.1 WordNet Queries

Analysis begins with a series of hypernym queries to WordNet. A hypernym query returns the parent concept of a word. Because words have multiple tenses there are multiple pathways from leave to root. To catch all paths, the search queries must reach both upward for parent nodes and horizontally for sibling word tenses.

Nouns and verbs are organized based on their conceptual meaning rather than their syntactic qualities. Words and the concepts they represent are interchangeable. Nouns and verbs exist in synsets of conceptually similar words. Every synset has a 'head word' that typifies the concept. The hierarchy in WordNet depends on head words, they make it possible to refer to a synset without enumerating every member of the synset.

WordNet can be used as a free-standing application, compiled and run as terminal application or incorporated and called from an entirely different application. Similarly, WordNet can be queried through its application interface, its command line interface, or through function calls from unrelated applications.

Figure 7.7 demonstrates the typical results returned from a WordNet hypernym query. This query asked for the hypernyms of the noun, 'newspaper'. Four senses of newspaper are returned and the listing displays the hierarchy of subsuming synsets from each sense of the word 'newspaper' to the root, 'entity'. Each word sense refers to a distinct concept of newspaper. The first word in the synset is the head word for the synset. Figure 7.8 shows a graphical representation of the results in Figure 7.7.



```
Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun newspaper
4 senses of newspaper
Sense 1
newspaper, paper
       => press, public press
           => print media
               => medium
                   => instrumentality, instrumentation
                       => artifact, artefact
                           => whole, unit
                               => object, physical object
                                   => physical entity
                                       => entity
Sense 2
newspaper, paper, newspaper publisher
       => publisher, publishing house, publishing firm, publishing company
           => firm, house, business firm
               => business, concern, business concern, business organization, business organisation
                   => enterprise
                       => organization, organisation
                           => social group
                               => group, grouping
                                   => abstraction, abstract entity
                                       => entity
Sense 3
newspaper, paper
       => product, production
           => creation
               => artifact, artefact
                   => whole, unit
                       => object, physical object
                           => physical entity
                               => entity
Sense 4
newspaper, newsprint
       => paper
           => material, stuff
               => substance
                   => matter
                       => physical entity
                           => entity
                   => part, portion, component part, component, constituent
                       => relation
                           => abstraction, abstract entity
                               => entity
```

Figure 7.7 WordNet Query Result.

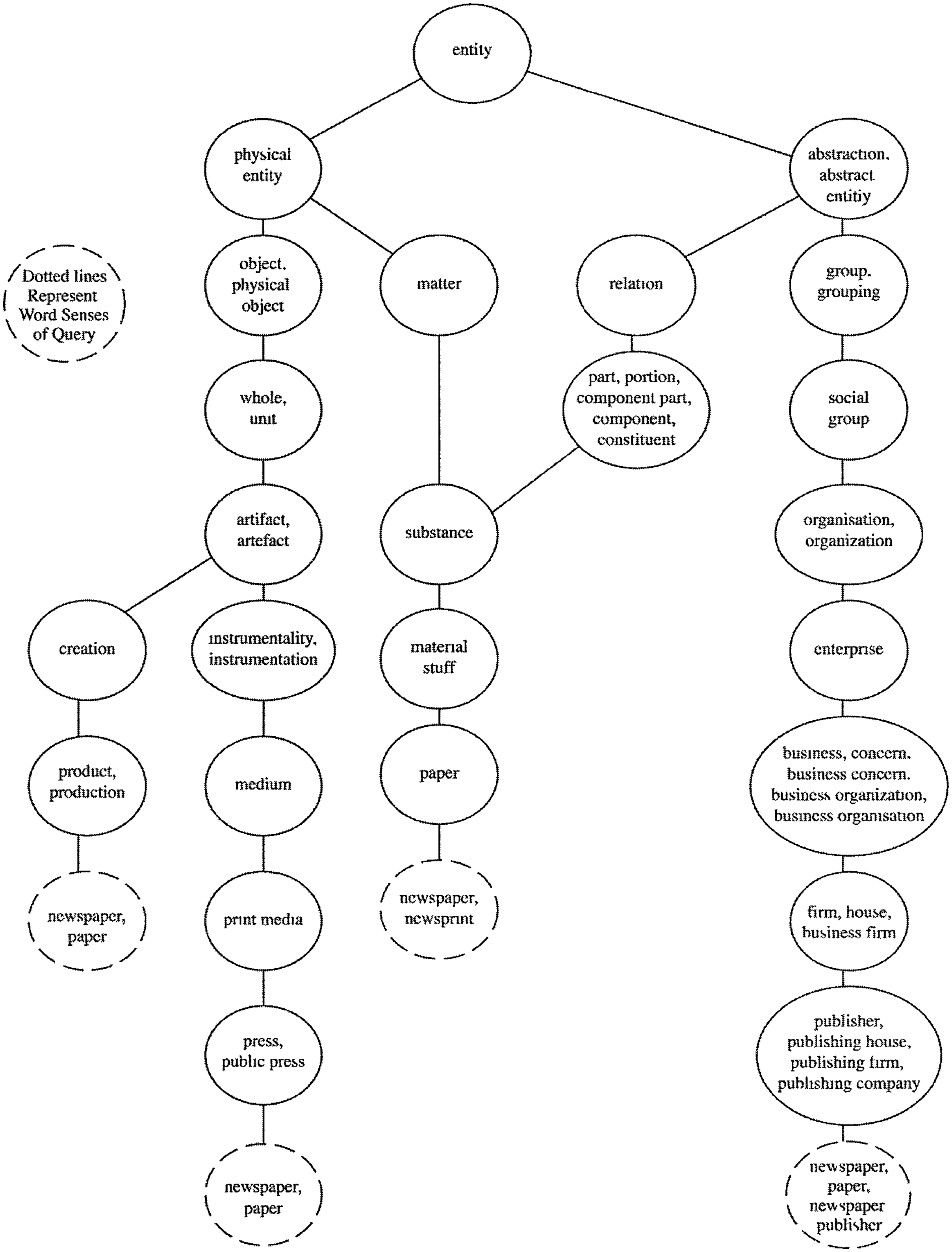

entity
physical entity
abstraction, abstract entitiy
Dotted lines Represent Word Senses of Query
object, physical object
matter
relation
group, grouping
whole, unit
part, portion, component part, component, constituent
social group
artifact, artefact
substance
organisation, organization
creation
instrumentality, instrumentation
material stuff
enterprise
product, production
medium
paper
business, concern, business concern, business organization, business organisation
newspaper, paper
print media
newspaper, newsprint
firm, house, business firm
press, public press
publisher, publishing house, publishing firm, publishing company
newspaper, paper
newspaper, paper, newspaper publisher

Figure 7.8 Graphical Representation of WordNet Query Result.

7.6.2 Edge Lists

One edge list is generated for each WordNet word sense discovered during the query process. Figure 7.9 demonstrates an example of an edge list, showing the path to the third tense of the noun 'speculation'. The edge list literally represents the physical location of the search word in WordNet's hierarchical structure. Any other search term that shares any edges with this edge list, will be somewhat similar conceptually.

```
entity01abstraction06relation01possession02assets01investment02speculation03
```

Figure 7.9 Noun Edge List.

Edge lists represents the exact coordinates from WordNet's root 'entity' to a given word sense. They are the reverse order concatenation of 'head-word' and word tense pairs that represent the path from a particular leaf word to root. Another way to look at the edge list in Figure 7.9 is to read it right to left. In this order, it shows the successive subsuming concepts separating a search term from the root.

Nouns in WordNet are grouped under the single root concept, 'entity01'. Verbs are grouped under 15 basic root concepts. The difference in hierarchical style, i.e. nouns stored in a single deep tree and verbs stored in several shallow trees, is reflected in the generated edge lists. Figure 7.10 shows how verb edge lists tend to be considerably shorter than noun edge lists.

```
make03create_by_mental_act01plan03
```

Figure 7.10 Verb Edge List.

Edge list coordinates designate the locations of words based on WordNet's semantic hierarchy. The parent sense of each word is the nearest subsuming concept. Moving up the tree, each parent concept represents a more general concept.

7.6.3 Edge lists and Similarity Detection

Edge lists can be used to detect similarity. Shared edges indicate similarity and distinct edges show dissimilarity. The last shared concept between two or more edges, is the nearest subsuming concept for all.

Consider the similarity of the words: hamburger, papaya, and macaroni. Lexically these words don't share any common traits, so morphological processing isn't likely to reveal anything interesting. Definitions aren't likely to help in assessing similarity, either. One is a fruit, one is a form of animal meat, and the other is a processed grain. Edge lists, however, reveal a commonality and show how these concepts are related.

```
entity01physical_entity01matter03solid01food02meat01beef02ground_beef01
entity01physical_entity01matter03substance07food01nutriment01dish02-
        snack_food01sandwich01hamburger01
entity01physical_entity01matter03solid01food02pasta02macaroni02
entity01physical_entity01causal_agent01person01adult01man01dandy01macaroni01
entity01physical_entity01object01whole02living_thing01organism01person01adult01-
        man01dandy01macaroni01
entity01physical_entity01causal_agent01person01male02man01dandy01macaroni01
entity01physical_entity01object01whole02living_thing01organism01person01male02-
        man01dandy01macaroni01
entity01physical_entity01object01whole02natural_object01plant_part01plant_organ01-
        reproductive_structure01fruit01edible_fruit01papaya02
entity01physical_entity01matter03solid01food02produce01edible_fruit01papaya02
entity01physical_entity01object01whole02living_thing01organism01plant02-
        vascular_plant01woody_plant01tree01angiospermous_tree01fruit_tree01papaya01
```

Figure 7.11 All Edge Lists.

Figure 7.11 shows the complete set of edge lists generated from WordNet for the nouns: hamburger, papaya and macaroni. A subset of the edge lists which share the longest prefixes shown in Figure 7.12. Finally, Figure 7.13 demonstrates graphically how edge lists can be used to discover similarity. The graph shows that the concepts are similar, they are in fact all subsumed by the second tense of the concept 'food'.

entity01physical_entity01matter03solid01food02meat01beef02ground_beef01
entity01physical_entity01matter03solid01food02pasta02macaroni02
entity01physical_entity01matter03solid01food02produce01edible_fruit01papaya02

Figure 7.12 Subset of Edge Lists.

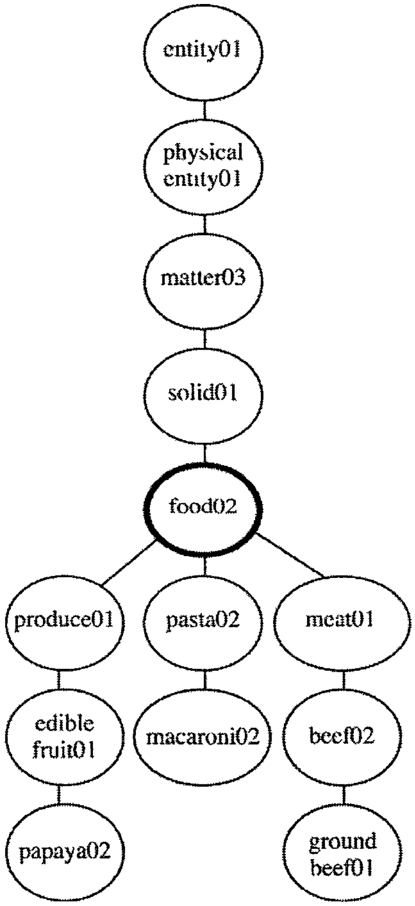


Figure 7.13 Subset of Edge Lists Represented as a Graph.

WordNet's conceptual hierarchy allows for the generation of semantic networks that capture the variety of meanings that a document may have. The hope is that by casting a wide conceptual net, similarity will be caught.

7.7 Similarity Comparison

After all of the edge lists for a document are identified, the EdgeNode list is augmented with the proper noun n-grams and the captured tagged strings. The EdgeNode list contains at this point, the complete feature set of a document.

7.7.1 Finite State Machines

For document comparison, the edge-list is used to create a finite state machine– reflecting the semantic network of possible meanings that the document may have. Each edge in the edge list becomes a state, each ending edge becomes an accepting state. The FSM is the structure that allows semantic distance to be measured.

Figure 7.14 shows how the EdgeNode list shown in Figure 7.11 would be translated into an FSM. This is a very simple example of three nouns. Documents from the data set generate semantic networks that are much larger. The larger the number of features, the more edge lists produced. The more edge-lists, the wider the semantic scope.

During comparisons, edges that end in accepting states are given higher total values and are seen as true matches. All other comparisons, are based on the ratio of the number of matching (or non-matching) edges to the total number of edges.

Similarity comparison is a measurement of the common features of a document minus the distinct features. To this end several methods of calculating similarity have been developed. The common goal is to measure similarity through semantic distance.

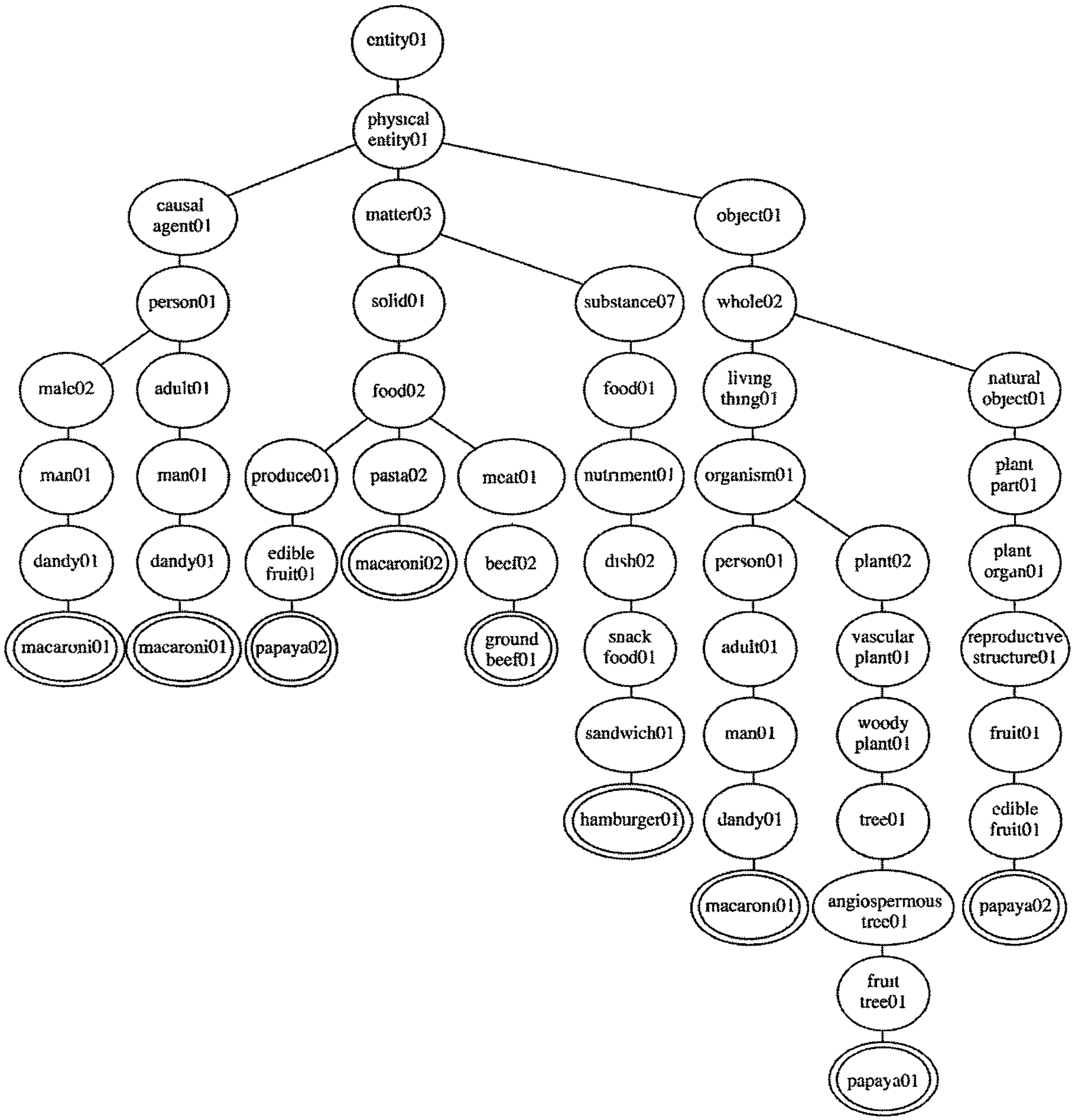


Figure 7.14 Finite State Machine Derived from Edge Lists of Figure 7.11.

7.8 Similarity Evaluation

During similarity evaluation, all of the documents in a data set are compared to each other. Every document's EdgeNode list is compared to every document's finite state machine. A document compared to itself will have perfect similarity. A document compared to any other document will have a score assigned that represents its similarity.

The results are displayed in a matrix where documents FSMs run vertically and EdgeNode lists run horizontally. Figure 7.15 demonstrates which document comparisons are made in each matrix cell. Similarity is asymmetrical, so there is no expectation that the matrix will mirror itself.

	Document 1	Document 2	Document 3	......	Document n
Document 1	FSM[1]EdgeList[1]	FSM[1]EdgeList[2]	FSM[1]EdgeList[3]	....	FSM[1]EdgeList[n]
Document 2	FSM[2]EdgeList[1]	FSM[2]EdgeList[2]	FSM[2]EdgeList[3]	....	FSM[2]EdgeList[n]
Document 3	FSM[3]EdgeList[1]	FSM[3]EdgeList[2]	FSM[3]EdgeList[3]	......	FSM[3]EdgeList[n]
....	....	....	......	......	....
Document n	FSM[n]EdgeList[1]	FSM[n]EdgeList[2]	FSM[n]EdgeList[3]	....	FSM[n]EdgeList[n]

Figure 7.15 Similarity Matrix Layout.

7.8.1 Contrast Model

The Contrast Model approach focuses on simple edge counting. Similarity is based on the total number of shared edges between two documents less their distinct edges. The resulting matrix stores, for each document pair, the sum of their shared features less the number of their distinct features.

7.8.2 Ratio Model

The algorithm is based on the Ratio Model. It counts the number of edges that two documents share. This is done by summing the number of matching path prefixes shared by one document's edge list and another document's FSM. The number of edges that distinguish the edge list of one document from the FSM of another are also counted. The resulting matrix for this algorithm, stores for each document pair, the sum of their shared features divided by the sum of their shared features plus their distinct features.

7.8.3 Pearson Product Moment Correlation Coefficient

This comparison is based on the Pearson Product Moment Correlation Coefficient, described in section 2.9.1. The algorithm is tightly related to the equation shown in Figure 2.8 and is fully described in section 8.3.3.

CHAPTER 8

IMPLEMENTATION

The first part of this chapter describes the data structures used to capture and preserve the features of each document. The second part explains which algorithms have been used to fill the structures and analyze the data. The third section describes the algorithms used in assessing the similarity between documents.

8.1 Data Structures

The process begins by reading in a set of documents into a single Corpora object, Figure 8.1. The Corpora object is the hardest working object in similarity comparison. It is responsible for reading the data set input, provides the framework for document comparison and holds the rule sets and lexicon used by all Document objects.

8.1.1 Corpora Objects

The Corpora class stores the data of each SGML document as a unique object in its Document array. It holds the document array; the dictionary of words and frequently associated part-of-speech tags–*lexTable*; two sets of rules used in part-of-speech tagging–the set of contextual rules, *cRuleList*, and the set of lexical rules, *lRuleList*; and the set of suffix removal rules, *sRuleList*, used in lexical normalization. These elements are created once in the Corpora class and used by every Document object.

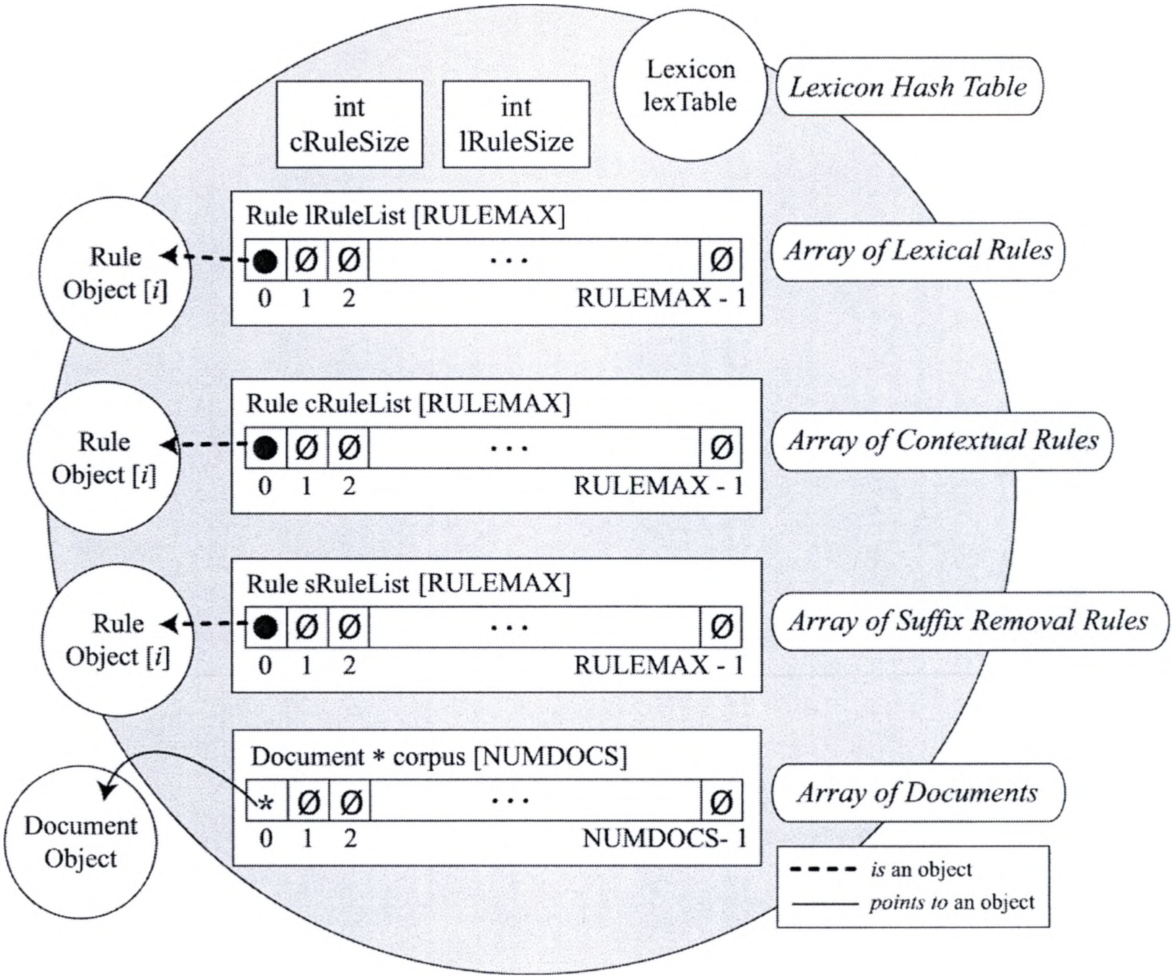


Figure 8.1 Corpora Object Diagram.

8.1.2 Document Objects

Document objects are used to preserve the relevant data found in SGML documents, Figure 8.2. This includes the storage of tagged data–stored as strings; the body copy of an article–stored in a syntax tree; the words of the body copy–stored with frequency counts in a string trie; the non-stoplist/non-proper noun words to be semantically analyzed–stored as word/part-of-speech pairs in a linked list; the proper noun n-grams stored in a unique linked list used during feature analysis; edge lists for all analyzed words designating the semantic path(s) for each word; and the FSM generated from the complete EdgeNode list, which holds a document's complete set of features. In short, Document objects contain the all of the structures necessary for deriving a complete set of features from a document.

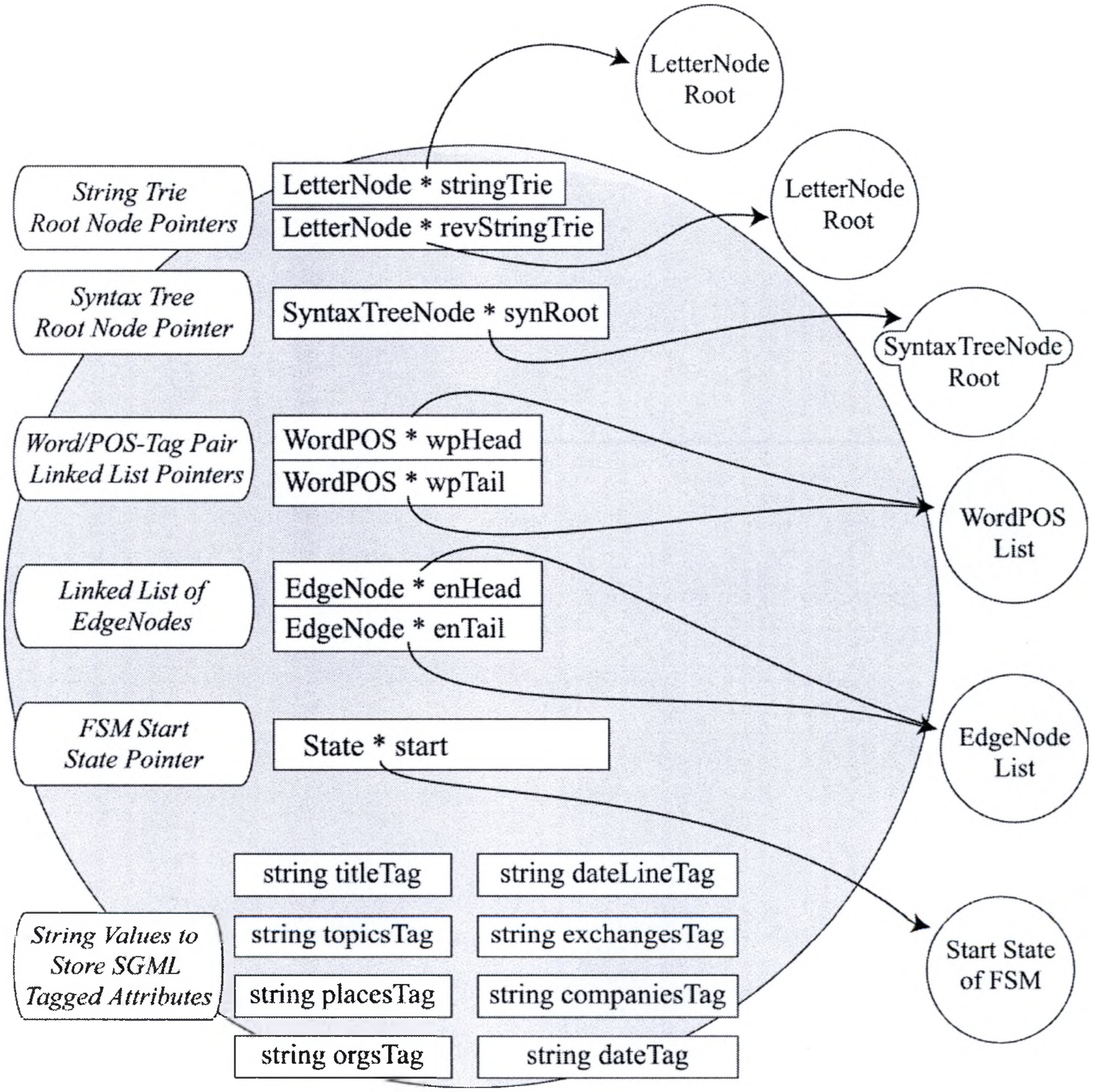


Figure 8.2 Document Object Diagram.

8.1.3 Unwrapping SGML Data

The first stage of processing retrieves data encapsulated in SGML tags from the beginning of the document until the BODY tag is reached. A parsing algorithm looks for attribute values within tags. If values are found they are stored in the current Document object as string values for their respective attributes. Tagged data encountered before the BODY tag of an article are used to extend a document's feature set.

The data within the BODY tag are read and stored in three different Document data structures. These structures are referenced by the pointers *stringTrie* and *revStringTrie*—which point to two distinct string trie structures, and *synRoot*—which points to a syntax tree. When the end of the BODY tag is reached, the Corpora object finishes the semantic anaylsis of the document and if the end of file marker has not been reached, it creates a new Document and the process repeats.

8.1.4 String Tries and LetterNode Objects

A string trie structure is used to quickly store and count the words in a document. Legal letters are stored in a LetterNode object-based string trie in the same order they are read in. Figure 8.3 demonstrates the composition of a LetterNode object.

Flags are used to indicate whether the original case of a letter was upper or lower; if a letter is the last letter of a word; if a word is a part of a larger proper noun n-gram; or if a word is part of a title.

Nodes flagged as the last letter of a word contain a pointer to a list of Location nodes and a count of how often the word appears. Location nodes store a set of coordinates identifying where each word occurs in the syntax tree. These nodes are used to store and retrieve words and their part-of-speech tags from the syntax tree structure.

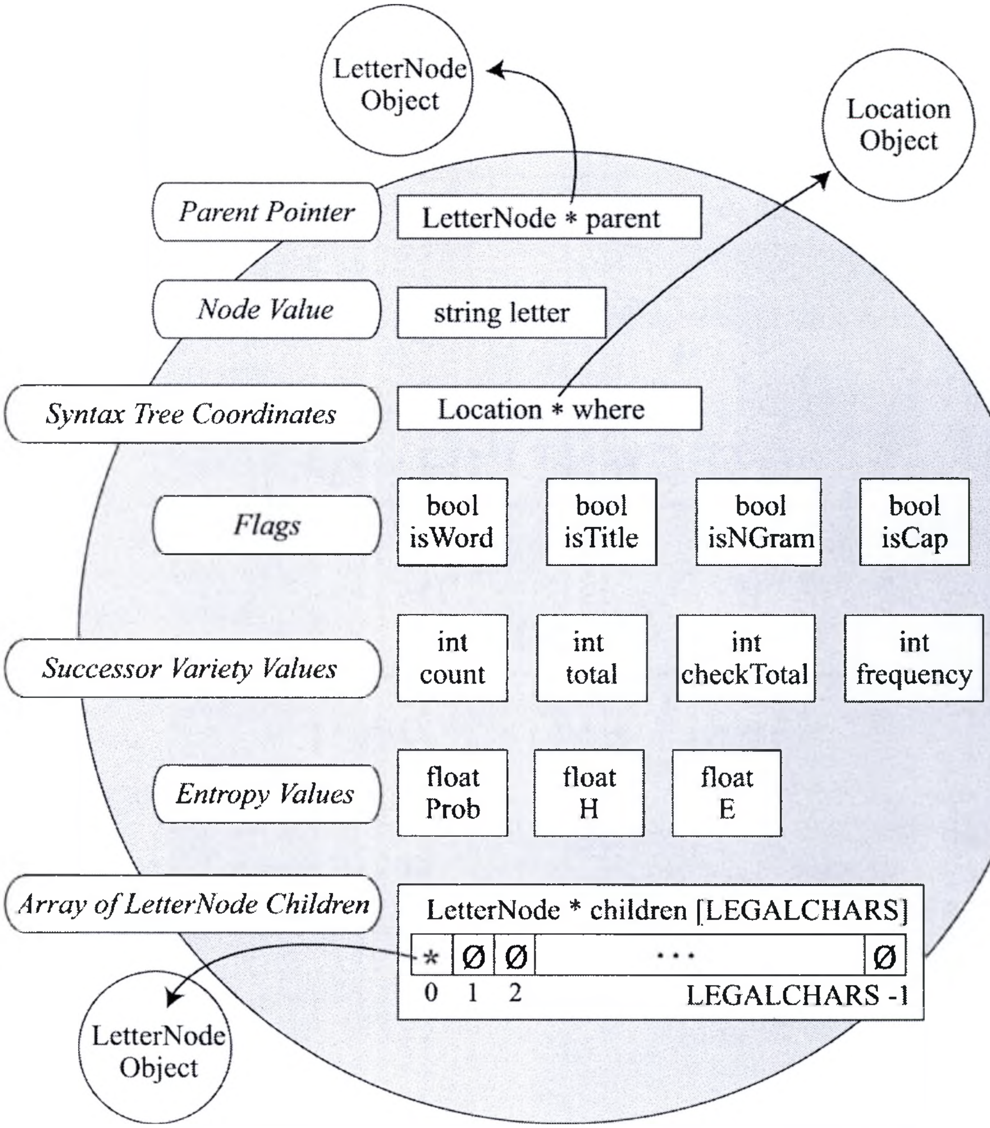


Figure 8.3 LetterNode Structure.

The difference between the two string tries, *stringTrie* and *revStringTrie*, is the order in which the letters are stored. The tree referenced by the *stringTrie* pointer stores words in order, as the letters are encountered. The tree referenced by *revStringTrie* stores words in reversed order–once an end of word is reached. The tree pointed to by *revStringTrie* is not used in the current implementation. The structure remains in place for the future development of prefix manipulation.

Using the theory behind Letter Successor Variety, described in chapter four, word stems are identifiable where branching of children is greater than one and an end of word flag is true. In order string trie structures are used to determine word stems–especially roots and suffixes.

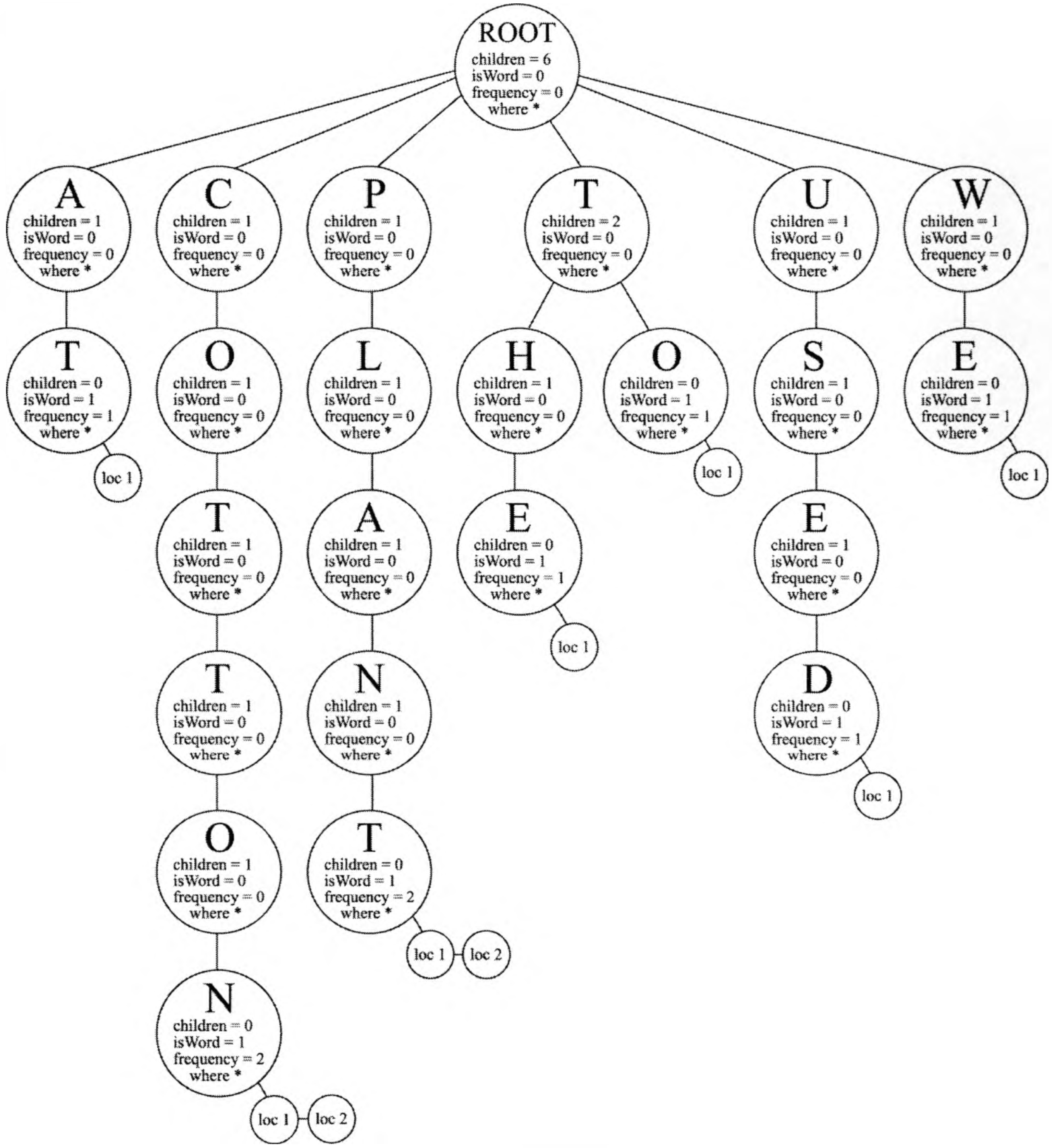


Figure 8.4 String Trie Structure.

Figure 8.4 shows how the sentence '*We used to plant cotton at the cotton plant.*' is stored in a string trie. Note the Location node lists pointed to by the *where* pointer. Location nodes are referenced from the last LetterNode of a word. They indicate the exact coordinates in the syntax tree where the word occurs. These coordinates are used to retrieve the part of speech tags for a word. Words with multiple frequencies have *where* lists with multiple nodes.

Reversed order tries are similarly used for determining segmentation between roots and their prefixes. Future development may include semantic analysis and categorization of words based on the meaning of their prefix. Because prefixes alter the semantics of words, this work focuses on suffix manipulation to change word tense.

8.1.5 Syntax Trees and SyntaxTreeNode Objects

To preserve the semantics of each document, words are added to the syntax tree a sentence at a time. Before being added to the syntax tree, each word has its part-of-speech identified and is transferred into a SyntaxTreeNode object, Figure 8.5.

Sentences are transformed from strings into sentence sub-trees of SyntaxTreeNode objects. In the syntax tree, the leaves are ordered sets of words with a sentence parent. Given a 'sentence-type' SyntaxTreeNode object with an array containing *n* indices, 0 through *n*-1, each index location will contain a pointer to each word in the sentence such that the original order is preserved.

Words are added to the syntax tree a sentence at a time. Memory is dynamically allocated as words, sentences and paragraphs are read in. Syntax trees are the structure used to remember the original syntax of each document. Figure 8.6 demonstrates the sub-tree generated for the sentence '*We used to pick cotton at the cotton plant* ' and Figure

8.7 shows how the sentence sub-tree would be incorporated into the larger syntax tree structure.

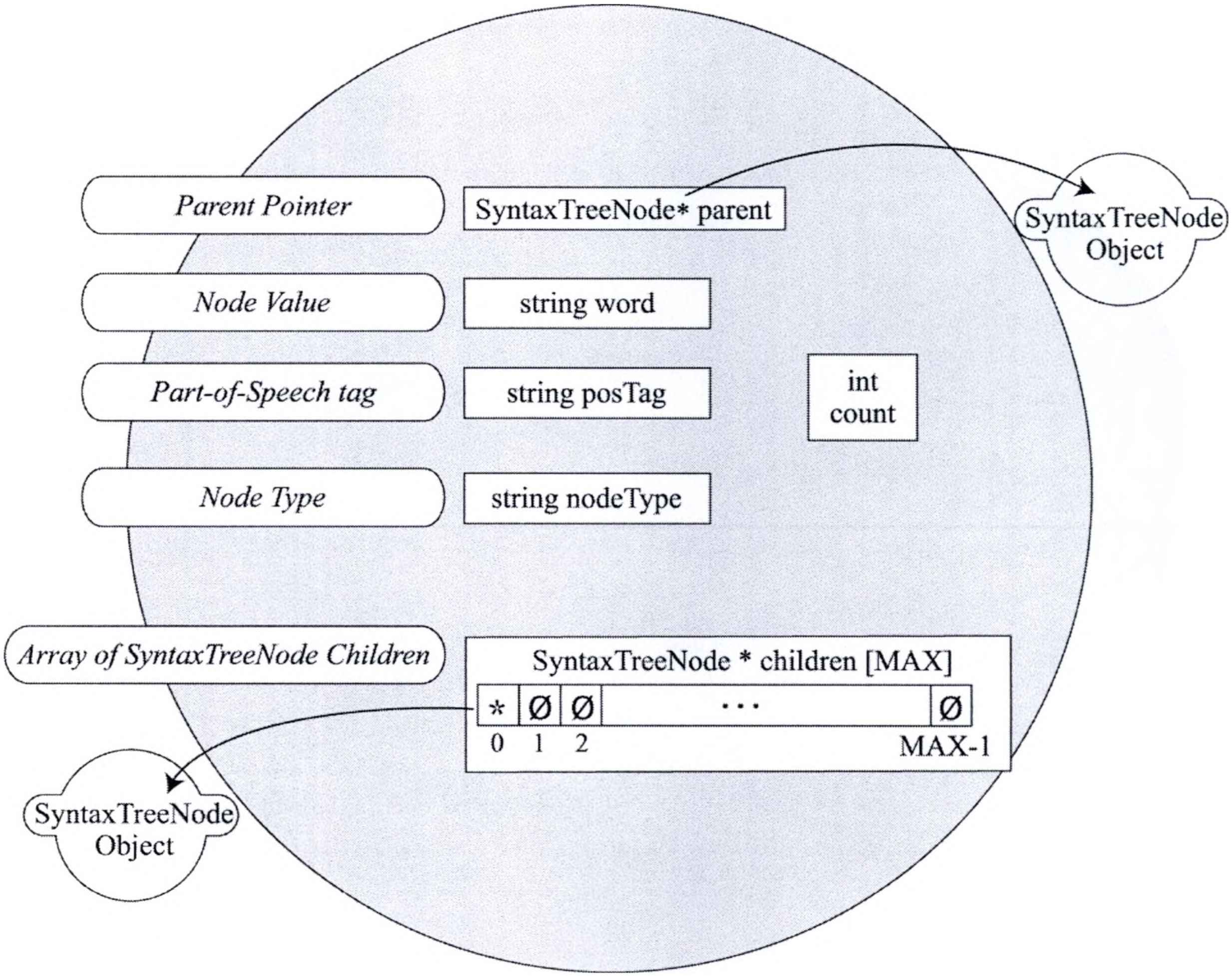


Figure 8.5 SyntaxTreeNode Structure.

This hierarchy continues upward. Sentences are stored under their paragraph parent, and paragraphs under their root parent, all in the order that they appear in the text. This hierarchy of arrays allows the part-of-speech tag for any word to be accessed and modified through lexical coordinates of the form (paragraph, sentence, word).

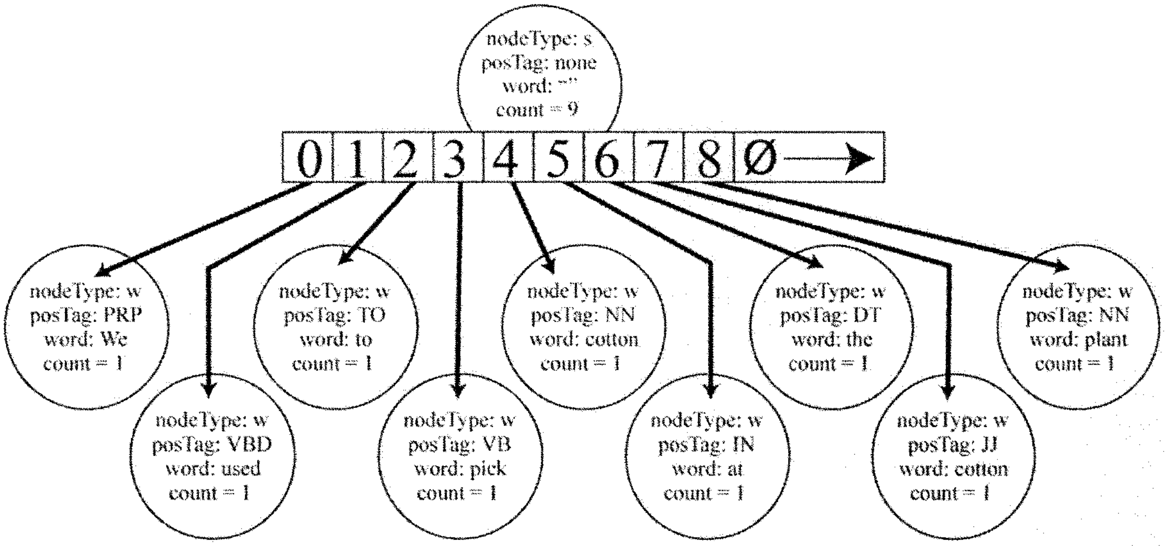


Figure 8.6 Syntax Tree Example Sentence Sub-Tree.

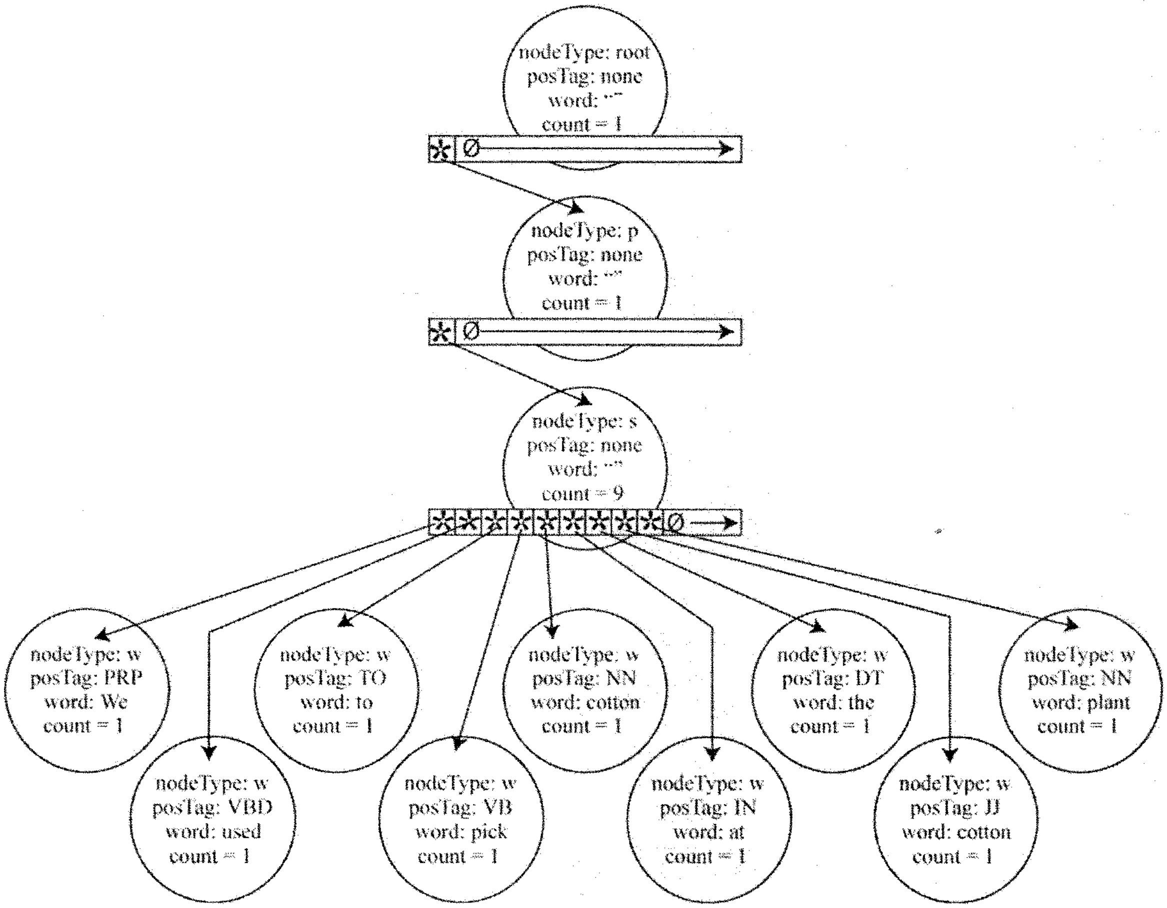


Figure 8.7 Syntax Tree Example of Total Structure.

8.1.6 Lexicon

The first stage of tagging assigns a part-of-speech to a word by lexical look-up in *lexTable*. The lexicon, *lexTable,* is a hash table that holds the dictionary included with Brill's tagger code. There are 93,696 word/tag pairs in the lexicon provided with Brill's code. To accommodate these elements and allow the addition of new words, the hash table holds 300,000 Entry structures. Entries are simple structure that encapsulates the data needed to store and retrieve words from the lexicon–namely the word, it's possible tags and its hash key. If a word doesn't appear in *lexTable*, then the word is given a default NN tag (NNP if the first letter is capitalized, NNS if the word is plural, NNPS is the word is plural and the first letter is capitalized).

8.1.7 Rule Objects

During the tagging process, the lexical and contextual properties of each word are examined using Brill's rules to form a best guess at the appropriate part-of-speech tag for a word. Brill's rule sets are read in and stored as arrays of Rule objects, Figure 8.8. Brill's contextual rules are stored in Corpora as the *cRuleList* array and his lexical rules are stored as the *lRuleList* array. Similarly, the suffix rules used to remove suffixes during lexical normalization are stored in Corpora as the *sRuleList* array. These rules also use Rule objects to store each rule.

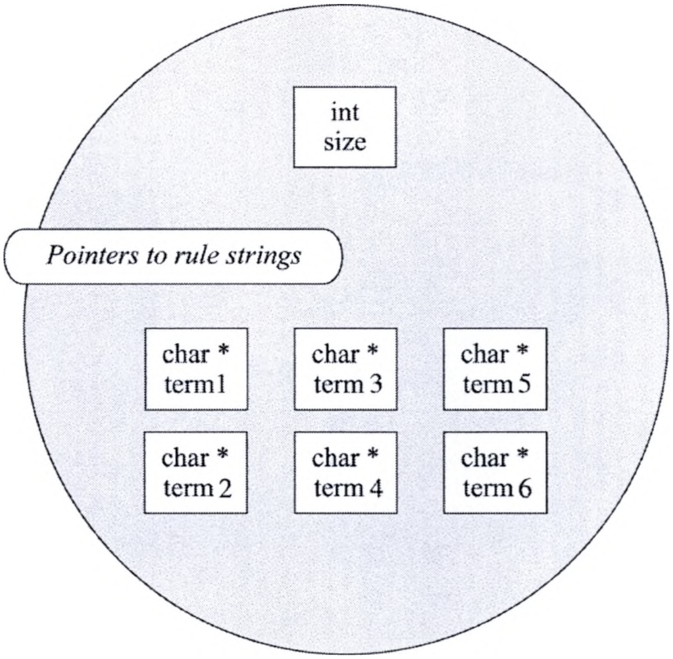


Figure 8.8 Rule Structure.

Rule objects have been custom made to fit the characteristics of Brill's rule sets. They have six character string pointers because Brill's rules have six terms. The functions associated with Rule objects are extremely simple, concerned only with setting and retrieving data.

The first stage of tagging applies each of lexical rules in the *lRuleList* array to a word and forms a guess at what the part of speech tag should be, based on the morphological properties of the word. The second stage of tagging contextually refines the initial guess by applying each of the contextual rules in the *cRuleList* array to the word. This set of rules considers not only the word and tag in question, but also the tags of the preceding three words and the subsequent three words.

At the end of this process, all of the words in a sentence are tagged and the word with its best-guess part-of-speech tag is added to the lexicon. Adding words to the lexicon allows the program to improve its default value guess stage by increasing the number of known words for future queries.

8.1.8 WordPOS and ProperNoun Objects

After the body of a document is read, the *stringTrie* is examined. The string trie holds all of the words in the document, including proper noun n-grams. In a depth-first traversal, words are retrieved and separated into two structures based on whether they're suited for semantic analysis, or not.

Proper noun n-grams are transferred into ProperNoun node structures, Figure 8.9. These nodes are linked into the Document class's ProperNoun list. Proper nouns are not suitable for semantic analysis and so they are set aside.

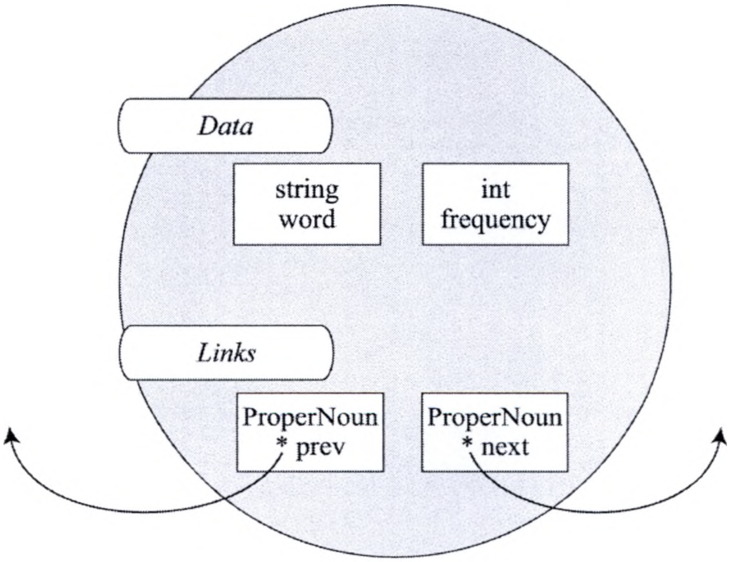


Figure 8.9 ProperNoun Structure.

All other words require more processing to determine whether they are fit for analysis. Before they are transferred into WordPOS structures they are processed to remove suffixes and trailing punctuation. Once they are in their root form, they are checked against stoplist words. Words that remain are transferred into WordPOS node structures, Figure 8.10.

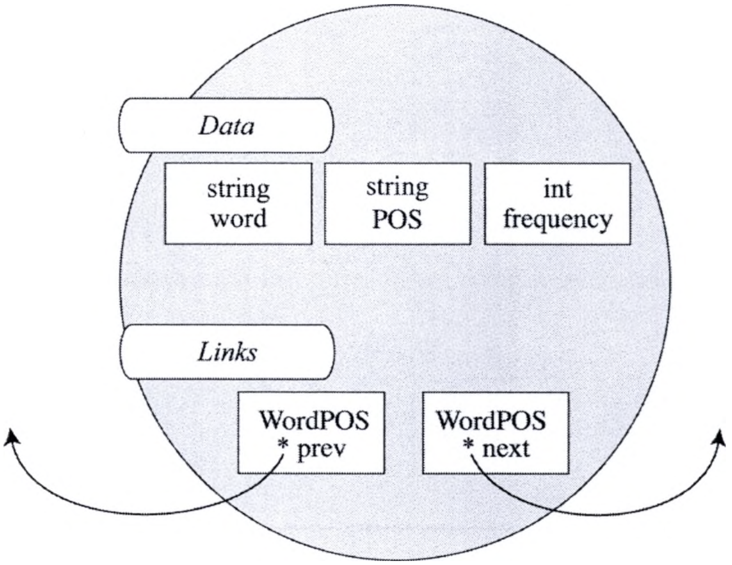


Figure 8.10 WordPOS Structure.

Words in the WordPOS list are grouped based on their part-of-speech tags. Their frequency conveys the number of times the word appears given a particular part of speech. This structure allows the verb 'plant' to be analyzed differently from the noun 'plant'. WordPOS nodes are linked into the Document class's WordPos list. Once the string trie traversal is complete, the elements in the WordPos list are processed for WordNet analysis.

8.1.9 EdgeNode Objects

Edge lists returned from WordNet analysis are stored in EdgeNode structures, Figure 8.11. Each element in the WordPOS list is transformed into semantic paths, or edge lists through a series of calls to WordNet. The number of edge lists generated for each word is dependent on the number of senses the word has. EdgeNodes combine an edge list for a particular word sense and its frequency of occurrence in the original

Document. Each EdgeNode structure contains a distinct edge list and is a unique feature of the document.

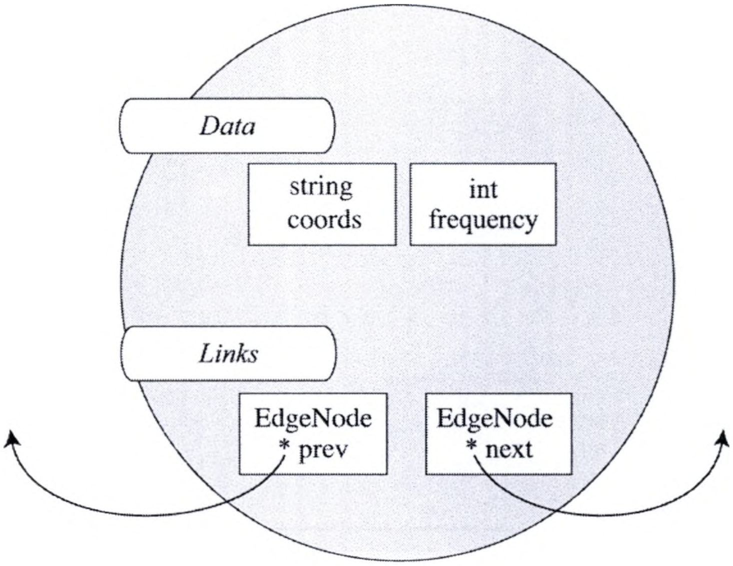


Figure 8.11 EdgeNode Structure.

After all of the WordPOS elements have been transformed into edge lists, all of the elements in the ProperNoun list are transformed into EdgeNodes and are added to the EdgeNode list. Finally, all of the captured SGML tagged data, preserved as strings in the Document class, are transformed into EdgeNodes and added to the EdgeNode list. Once these three types of features have been brought together, the documents feature set is complete.

8.1.10 State Objects

State objects are used to construct the finite state machine for the document, Figure 8.12. Each EdgeNode edge list is decomposed into a series of states. The last state of every edge list is designated as an accepting state. When edge lists are compared to the FSM, accepting state matches are scored higher non-accepting state matches- particularly on subsuming states that mark the end of shared features.

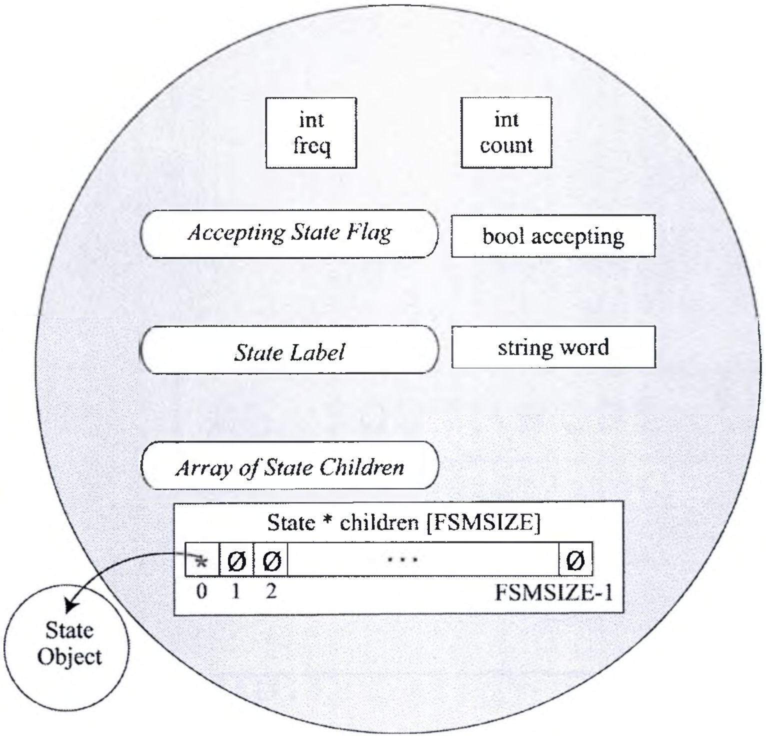


Figure 8.12 State Structure.

8.2 Implementation Algorithms

8.2.1 Lexicon

The lexicon, used to house the dictionary of words known to Document objects, is implemented as a hash table, in the Corpora class. The data source is the lexicon provided with Eric Brill's Part of Speech Tagger code. This lexicon Values in *lexTable* are hashed based on the hash key function shown in Figure 8.13.

```
unsigned long Lexicon :: hashkey ( char * str )
{   unsigned long hash = 63037;
    int c;
    while ( c = (int) * str++ )
        hash = (hash * 37 ) + hash + c;
    return hash;  }
```

Figure 8.13 Hash Function.

This key was chosen as it had the best '*spreadability*'. Out of 300,000 locations in the hash table, 80,007 are used. 12,229 locations store more than one location, of which only 1330 store more than two.

The hash function uses star primes which are of the form: $6n(n-1) + 1$. In this function, the *hash* variable is a star prime such that n equals 103. Using star primes the 93,697 elements in the lexicon are spread over 80,007 locations. Figure 8.14 shows a comparison of how the type of prime number affected hash table location hits.

Type of Prime	Locations Hit	%age Unique Hits	%age >Two-Hits
Star	80,007	85%	13.0%
Lucky	79,958	85%	13.1%
Mersenne	48,615	49%	18.3%
Padovan	79,705	85%	13.3%

Figure 8.14 Prime Comparison.

To further randomize the values, the hash key includes the integer values of each letter in an entry word. The end key is the result of the summation of the current *hash* value multiplied by a second star prime (37) added to *hash* plus the integer value of each letter of the string.

8.2.2 Creating String Tries

Lower case letters that are accepted by Corpora must be translated to uppercase letters before being added to the string trie. Only uppercase alphabetic letters exist in the trie. This is done to reduce the overall complexity. Without this step, the same word could be hung in multiple ways in the trie. This would result in redundancy and inaccurate counts.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

U	V	W	X	Y	Z	0	1	2	3	4	5	6	7	8	9	'	-	.	#
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39

Figure 8.15 LetterNode Index Map.

String tries are created through the *hanging* of legal characters, one-at-a-time, as they are read in. A map displaying the index positions of legal characters is shown in Figure 8.15. Figure 8.16 demonstrates how the letters for the word "*Fire*" are hung in a string trie. This image highlights how the indices are used to navigate a path from the root of the trie to letter 'R".

If the word "*Fire*" appeared in text, the character after 'e' may be a space, hypen or a period. By looking a character ahead, spaces, hyphens and periods are used to signal the end of a word. An end-of-word flag indicates three things. The first is that the current letter is the last letter of a word and the *isWord* flag for this node must be set. As this flag is set, the frequency is incremented. The second is that if the next character is a space, then the LetterNode pointer used to navigate the trie, must return to the ROOT. The third action item is to increment the ROOT's count of total words in the trie, for verification purposes.

Special cases override the return to ROOT. These cases require that the current word is a proper noun, indicated by an initial cap. This case causes the LetterNode pointer to sit tight and wait for the next look-ahead to make its decision of whether to hang a space and continue down the tree or return to ROOT. If the character after the space is a capital letter, the pointer adds a space, moves ahead and prepares to hang the next proper noun word after the last. If the character after the space is a lowercase letter, number or symbol the pointer returns to ROOT.

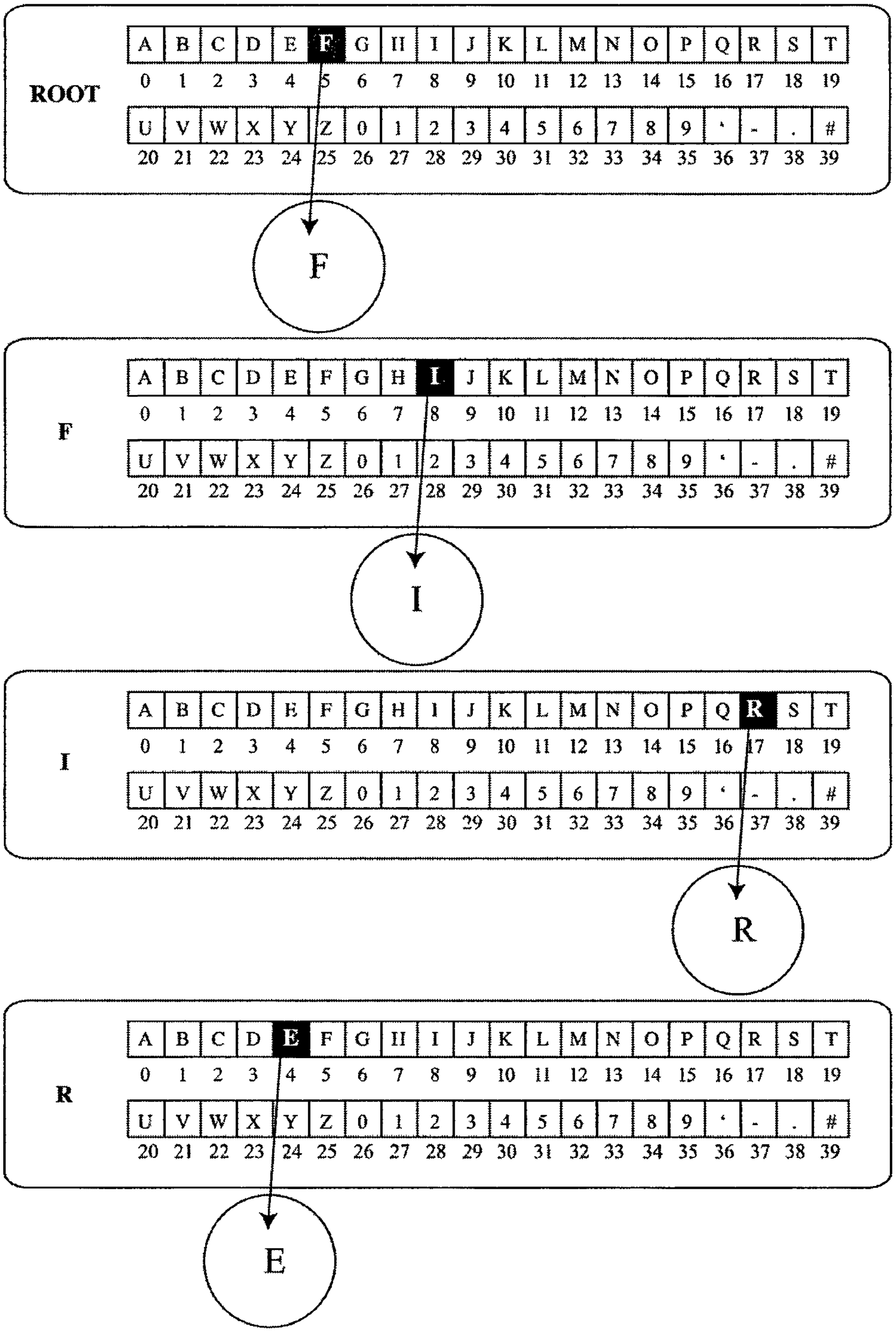


Figure 8.16 String Trie Index Navigation.

When a letter is to be added, the index where it should be placed is queried. If the letter is already there, the pointer simply moves to the index location, otherwise a new LetterNode is allocated and linked from the index location. Letters are added in this fashion until the end of the document has been reached.

After the document has been read, a depth first traversal of the trie generates a complete listing of all words in the document in alphabetical order. The listing provides more than just the words, it also includes their respective frequencies and a pointer to a list of Location nodes that indicate the exact position in the syntax tree where each word's part-of-speech can be found.

8.2.3 Part of Speech Tagging

Words are collected into sentences. When look ahead detects a period, a flag is set that triggers the processing of the sentence. The function *tagSentence*() is called with the sentence passed as the input parameter.

The first thing that occurs in *tagSentence*() is the allocation of the SyntaxTreeNode *sentence*. This node functions as the sub-tree root. Once words receive part-of-speech tags, they are hung in their appropriate position in the sub-tree's array.

The sentence is broken ınto ıt constituent words and the first step of Brill's tagging algorithm is performed. Individual words are looked up in the lexicon. If the word is found in the lexicon and a tag is assigned, the lexical rules are not applied to the word. If, however, the word is not found, the lexical rules are used to make a preliminary guess at the words part of speech. The function that implements Brills lexical rules, eval() is called. Because lexical rules look at the words immediately preceding and following a word, three parameters are passed–the word, the previous word and the next word.

Eval() assigns the appropriate noun tag based on the morphological properties of the word. Each rule in the rule list is fired to test if changing certain properties of the word cause it to be found in the lexicon. For example, a rule may test if the word has a particular suffix, if the word does then the rule instructs the function how to remove the suffix and then try to find the word in the lexicon again. Some rules add or remove suffixes and search for the word again. Others add or remove prefixes and search again for the word. Still others look at the preceding or next word's tag and make revisions to the words tag based on local context. At the end of *eval*(), the word and the best guess at its tag is added to the lexicon to improve the efficiency tagging in the future.

The best guess is returned from *eval*(). This tag is used as the preliminary tag for the word. The word and tag are transferred into SyntaxTree nodes and hung in their sentence sub-tree in the order they appeared in the original document. The entire sentence is returned as a pointer to the sentence sub-tree's root node.

This pointer is then passed as the parameter to the function that implements Brill's contextual rules. *ContextRefine*() looks at the three words and their part-of-speech tags preceding a word, the word and its tag, and the three words and their tags that follow. In a process very much like *eval*() the rules are applied to the words in the sentence. These words consider the tags surrounding the word and if tags meet requirements, the tag of the word is changed. These rules test a number of conditions looking at one, two, and three tags before, surrounding tags, the word and the second tag after, left bigrams right bigrams, and more.

ContextRefine() modifies the sentence via the pointer that's passed to it. After the context has refined all of the tags, the sentence is linked to the appropriate paragraph node, reflecting the position of the sentence in the original sentence.

8.2.4 Creating Syntax Trees

The syntax tree is generated a sentence at a time following the process of part-of-speech tagging. Each sentence of the tree is created as a sub-tree during the first phase of tagging. The position of where the sentence is hung is maintained through a sentence index counter.

Every time a sentence is identified, the counter is incremented. New paragraphs in Reuters documents are indicated by four successive spaces. When a sentence is flagged, a count of the trailing spaces is triggered. If this count equals four, the paragraph counter is incremented and the sentence and word count are reset to zero. This process continues until the end of the document.

8.2.5 Retrieving Part of Speech Tags

The first phase of lexical normalization requires the part-of-speech tags to be collected for each word. This is done during the traversal of the strıng trie that retrieves a listing of words, frequencies and locations for the document. The traversal starts at the root of the string trie and makes recursive calls to each child, in order. The first LetterNode considered is the bottom-most left node in the tree.

As the calls are processed, each LetterNode is tested to see if its *isWord* flag is set. If the flag is set, a Location pointer is directed to this node's *where* list. Each Location node's coordinates are used in a query to the syntax tree of the document and the part of speech tag for the word, given the location, ıs added to a temporary location string. While unexamined locations exist in the *where* list, locations are queried in the syntax tree and part of speech tags are added to the temporary location string.

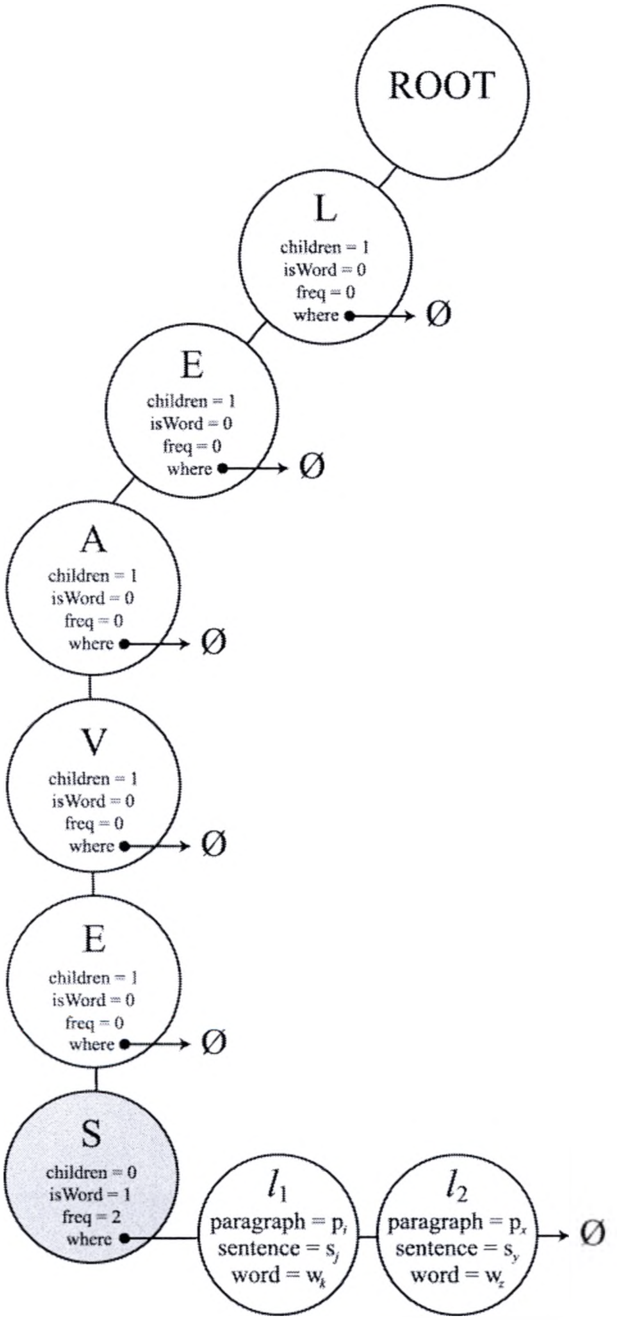


Figure 8.17 Location Nodes Used for Part-of-Speech Retrieval.

Figure 8.17 visualizes Location nodes stored from a LetterNode whose *isWord* flag is set. In this example, the word 'LEAVES' appears in two locations. The first location appears at coordinates (p_i, s_j, w_k) and the second location appears at coordinates (p_x, s_y, w_z). The coordinates (paragraph, sentence, word) relate to the index position of the paragraph held by the ROOT node, the index position in sentence held by the paragraph node, and the index position of the word held by the sentence node.

8.2.6 Unloading the String Trie

After the word's locations have been identified, the word is tested to see if it's a proper noun. If it is, the word is filtered through *strongWord*(). *StrongWord*() is a function that returns true if a word is not a number or a stoplist word. If a word is found to be a strong word, the word and its frequency are sent to *addToPN*() for further processing.

AddToPN() removes possessive endings ('s) and periods from the end of words. It transforms the word into a ProperNoun node and adds the ProperNoun node to the list of proper nouns. It then transforms the word into an EdgeNode node and adds the proper noun to the EdgeNode list. Proper nouns do not require their tags to be specified, as they distinguished by being in all uppercase letters.

If the word is not a proper noun, it is tested to see if it's a strong word. If it's strong, it's sent to the function *divideTags*() for further processing. These words are preprocessed for semantic analysis and used in the creation of the WordPOS list.

8.2.7 Creating WordPOS Lists

DivideTags() is used to create the WordPOS list. It receives as parameters a word and a tag list. It looks at the tag list and creates homogenously tagged WordPOS nodes. If a tag list contains different tags for the same word, the tags and are divided into separate categories and a unique WordPOS node is created for each type of tag. Figure 7.4

demonstrates a word with multiple tag types and Figure 7.5 shows how WordPOS nodes are used to separate words based on their tags.

Before words are transformed into WordPOS objects, they are filtered through the functions *expandContractions*() and *removeSuffix*(). Reuters-21578 documents use abbreviations for frequently occurring words. These abbreviations are expanded to their full word value in *expandContractions*(). *RemoveSuffix*() uses the suffix removal rules to determine if a word has a suffix. If a suffix is discovered, the rules on how to remove the suffix given the morphology of the word are examined. If a match is made between how to remove the suffix and the morphology of the word, the suffix is removed.

Each divided tag is combined with the processed word and transformed into a WordPOS contender node. Because words have their suffix removed, a search of the WordPOS list is required before adding, to see if the word is already there. If the word is there, a check is conducted to see if the word is there *with* the current tag. If a match is found, the existing node in the WordPOS list has its frequency incremented the amount of the WordPOS contender's frequency and the WordPOS contender is deleted. If a match is not found on the word or if a match is found on the word but not the tag, the WordPOS contender is added to the WordPOS list after the last matching word.

8.2.8 Creating Edge Lists

After the process of lexical normalization is complete, semantic analysis begins. The Corpora class guides semantic analysis by extracting and passing the word, part-of-speech tag and frequency from each WordPOS node to the function*findtheinfo_aux*().

Findtheinfo_aux() is a wrapper function that generates edge lists through recursive calls to WordNet. This function makes the initial call to WordNet's *findtheinfo*() function and then uses the returned data structure to navigate a path horizontally to include alternative senses and a path upward from word sense to root.

The WordNet hierarchy is navigated through the use of WordNet's Synset objects, Figure 8.18. These objects are returned from WordNet queries and encapsulate everything you could want to know about a word. This includes the address where they're stored in memory, what they're called, what they're definition is, what synsets they are similar to, what synsets include similarly spelled words, and more.

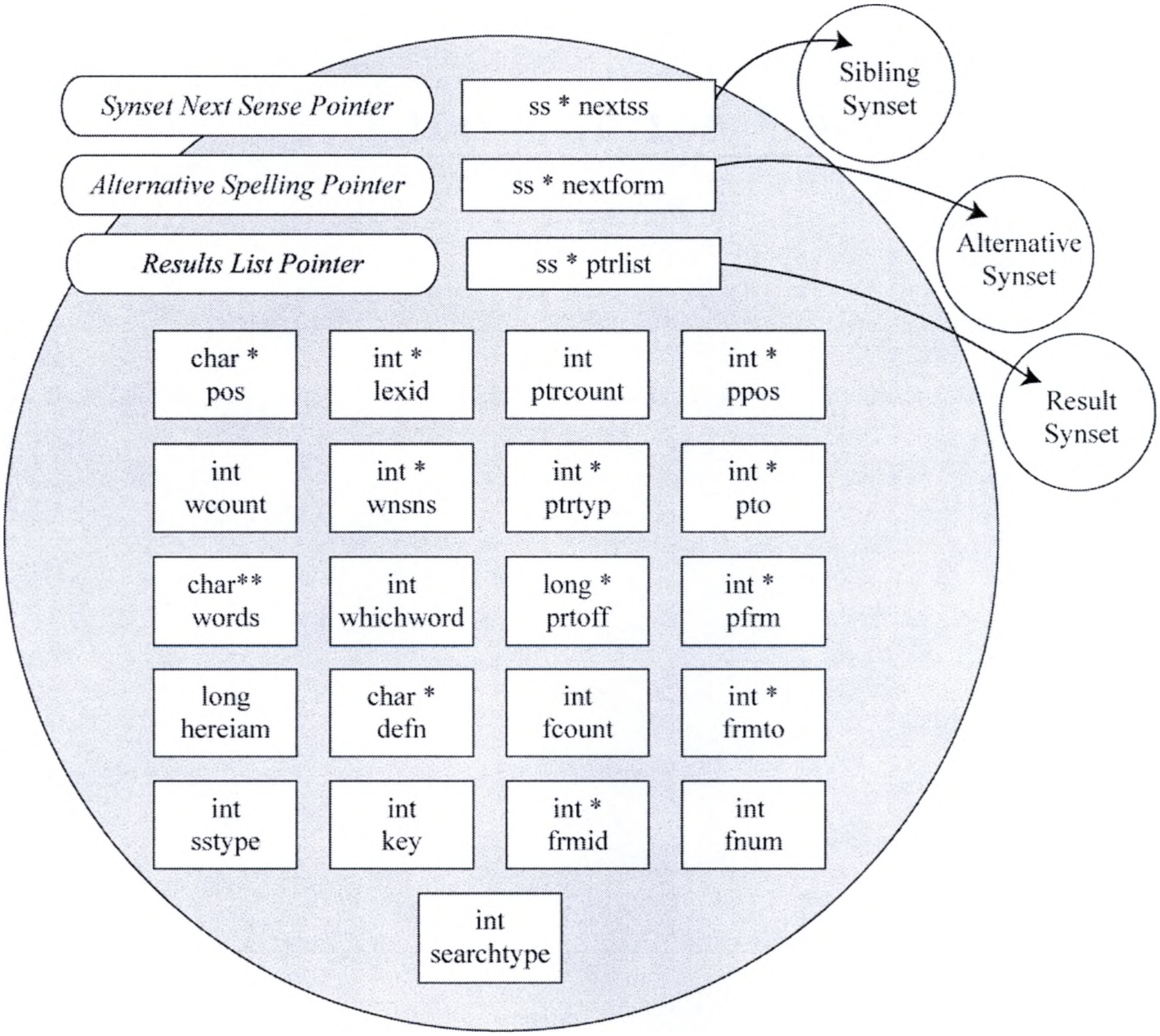


Figure 8.18 Synset Objects.

```
void Document :: findtheinfo_aux (string coords, char* word, int sense, int pos, int searchtype)
{   SynsetPtr ptr = findtheinfo_ds(word, pos, searchtype, sense);
    SynsetPtr aptr;
    bool done = false;
    char * p = word;
     int ws;
    while(!done && coords.size()<200)
   {    p=*ptr->words;
        ws = *ptr->wnsns;
        if(ptr->nextss)
       {    aptr = ptr->nextss;
            while(aptr)
           {    findtheinfo_aux(coords, *aptr->words, *aptr->wnsns, pos, searchtype);
                aptr=aptr->nextss;  } }
        coords = ws + p + coords;
        ptr =  findtheinfo_ds(p, pos, searchtype, ws);
        if(ptr->ptrlist)
            ptr = ptr->ptrlist;
        else
            done = true;  }
    if(coords.size()<200 && islower( coords[0]))
        addEdge(coords, f);
}
```

Figure 8.19 Semantic Analysis Function.

Figure 8.19 displays the code used in semantic analysis. The Synset pointer that is returned from the original WordNet query is used to create all of the edge lists covering all of the sense of each word. An outer loop searches upward for the next subsuming concept. An inner loop searches horizontally for next sense pointers. If a *nextss* pointer is found, a recursive call is made that searches upward from that point. Each time a subsuming concept is discovered in the outer loop, the inner loop looks for next sense pointers. Edge lists are the result of vertically mining the 'is-part-of' relationship and horizontally searching for *nextss* pointers.

Cycles do rarely occur in the WordNet hierarchy. To catch cycles, coordinate lists are required to have less than 200 characters. Coordinate lists with more characters are ignored. Approximately six out every thousand edge lists are ignored. Cycles were especially noticeable when adjectives were considered. The edge lists created from them were prone to cycle and lacked conceptual hierarchy. For these reasons, only nouns and verbs are currently considered during semantic analysis.

8.2.9 Creating the Finite State Machine

The finite state machine for a document is created after semantic analysis. The EdgeNode list is used as input. Each path is decomposed into a series of states. The FSM structure removes duplication and simplifies the EdgeNode list into a minimum structure representing all the states.

The type of path determines how the path is split. There are three types of EdgeNode: proper noun n-grams, edge lists returned from WordNet, or strings of captured data from SGML tags. These three type of edges are shown in Figure 8.20.

Proper Noun N-Grams:	Captain John Smith
Edge List:	entity01abstraction06relation01possession02assets01
SGML Tags:	usa

Figure 8.20 Three Types of Edge Lists.

Even though the types of EdgeNode lists vary, they are processed practically the same. Spaces are used to separate word chunks in proper noun n-grams and the words retrieved from SGML tags and sense numbers indicate the end of word chunks in EdgeNode lists.

A State pointer is directed at the *start* state. As each word chunk is encountered a search looks to see if the word chunk is already in the FSM. If it is, the pointer moves to the state, otherwise a new state is added with the word chunk stored as the state's label.

This continues until the end of the EdgeNode list is reached. The last element is processed the same way–except its state is flagged as an accepting state and the pointer is returned to the *start* state. Once all of the edge lists have been read in, the FSM is complete.

8.2.10 EdgeNode List Refinement

The FSM structure that flattens the EdgeNode list is used to create a minimized EdgeNode list to be used as an alternative approach in document comparison. Because the intersection of EdgeNode lists and FSMs are not equivalent, i.e. A ∩ B does not equal B ∩ A. A depth first traversal of each FSM is used to concatenate states from accepting states to root. This abridged EdgeNode list is referenced as an EdgeNode *featureSet*.

8.3 Comparison Algorithms

Comparison begins after all of the documents have been parsed, analyzed and have complete EdgeNode lists and finite state machines. Comparison algorithms identify similarity using ratio models, product moment correlation, and edge counts. Each examines a distinct approach to measuring similarity. The algorithms for comparison are presented in this section. Evaluationd of the algorithms are presented the next chapter.

8.3.1 Contrast Model

Semantic distance is the simplest measurement of similarity. The algorithm implementing this measurement is based on the Contrast Model, introduced in chapter two. This algorithm counts the number of common edges shared between two documents and the number of distinct edges in a function called *intersectionDifference()*. The pseudo-code for function is provided in Figure 8.21.

This function accepts two index values: *i* and *j*. These indices are used to query two objects in Corpora's Document array. The EdgeNode list in document *j* is compared to the FSM located in document *i*.

```
Psuedo-code for intersectionDifference(int i, int j)
{   State * ptr = sptr=getStart(i)
    EdgeNode *eptr = getFirstEdge(j)
    while( eptr )
    {   while ( eptr->isEdgeLeft() )
        {   temp = breakOffState()
            if(match)
            {   x= isChild(sptr, temp)
                if(x)
                {   sptr= getChild(x)
                    same++ }
                else
                {   match = false
                    different++     }   }
            else
                different++ }
        sumSim +=same
        sumDif +=different
        same = different = 0
        sptr=getStart(i)
        match = true
        eptr = getNextEdge() }
    if(sumSim+sumDif > 0)
    {   intersectionMatrix[i][j]=sumSim
        differenceMatrix[i][j]=sumDif }
}
```

Figure 8.21 Psuedo-Code for IntersectionDifference().

The comparison breaks apart each EdgeNode entry in Document *j* into its constituent states, one piece at a time. Each state is traced through Document *i*'s FSM. A State pointer starts by referencing Document *i*'s *start* state.

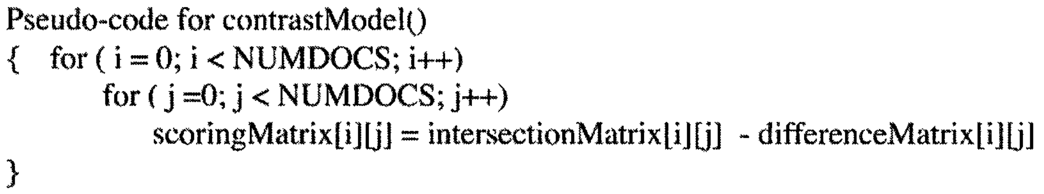
Every co-occurrence of EdgeNode state and FSM state causes the *same* counter to be incremented; sets the *match* flag; moves the State pointer to the matching reachable state; and the process repeats with the breaking off the next element from Document *j*'s EdgeNode edge list, if one exists.

If a state from the EdgeNode edge list is *not* found in the FSM, the *match* flag is turned off and the *difference* counter is incremented once for every additional state in the EdgeNode edge list.

At the end of each EdgeNode edge list, the values in *sumSim* and *sumDif* are incremented the amount held by the variables *same* and *different*. *SumSim* and *sumDif* track total edge similarity and edge difference, respectively.

After all of the EdgeNode elements are traced through Document *i*'s FSM the sum of all the same edges between Document *i* and Document *j*, *sumSim*, is stored in the matrix *intersectionMatrix* at position (*i*, *j*). Similarly, the sum of all the different edges, *sumDif*, is stored in *differenceMatrix* at position (*i*, *j*).

The second phase of this algorithm is implemented by a function called contrastModel(). The pseudo-code is provided in Figure 8.22. ContrastModel() uses a nested loop to fill the scoring matrix with the outcome of the equation.



```
Pseudo-code for contrastModel()
{   for ( i = 0; i < NUMDOCS; i++)
        for ( j =0; j < NUMDOCS; j++)
            scoringMatrix[i][j] = intersectionMatrix[i][j]  - differenceMatrix[i][j]
}
```

Figure 8.22 Pseudo-Code for ContrastModel().

Each position in the scoring matrix holds the result of the equation shown in Figure 8.23, note that SM = *scoringMatrix*, IM = *intersectionMatrix*, and DM = *differenceMatrix*.

$$SM[i][j] = IM[i][j] - DM[i][j]$$

Figure 8.23 Scoring Equation for the Contrast Model Calculation.

8.3.2 Ratio Model

This comparison technique implements the Ratio Model algorithm, shown in Figure 2.7. The code for the first phase begins with the same step as performed in the Contrast Model. It calls *intersectionDifference*(), shown in Figure 8.19. This function fills the intersectionMatrix with the sum of the matching path edges and differenceMatrix with the sum of the different edges.

The second phase of this algorithm is implemented by a call to the function, *ratioModel*(). This function uses a nested loop to fill each position in the *scoringMatrix*. The pseudo-code is shown n Figure 8.24.

```
Psuedo-code for ratioModel()
{  for (i = 0; i < NUMDOCS; i++)
       for ( j = 0; j < NUMDOCS; j++)
           if( i != j )
              scoringMatrix[i][j]= ( intersectionMatrix[i][j] /
              ( intersectionMatrix[i][j] + differenceMatrix[i][j] + differenceMatrix[j][i] ) )
}
```

Figure 8.24 Pseudo-Code for RatioModel().

Each position in the scoring matrix holds the result of the equation shown in Figure 8.25, note that SM = *scoringMatrix*, IM = *intersectionMatrix*, and DM = *differenceMatrix*.

$$SM[i][j] = \frac{IM[i][j]}{IM[i][j] + DM[i][j] + DM[j][i]}$$

Figure 8.25 Scoring Equation for Ratio Model Calculation.

8.3.3 Pearson Product Moment Correlation Coefficient

The Pearson Product Moment Correlation Coefficient is calculated based on the algorithm shown in Figure 2.8. The process begins by breaking an EdgeNode list from document *j* into its separate states. Each state is traced through the FSM in document *i.* Matches increment the *same* counter, once a difference is discovered the remaining states increment the *different* counter. Unique to this process, two arrays are used to track the ratio of same edges to total edges and different edges to total edges.

After the EdgeNode list has been examined the results are processed further. The mean, *sumMean*, is calculated dividing *sumSim*, the total similarity counter, by *edgeCounter*, the total number of edges. Similarly the mean, *difMean* is calculated by dividing *sumDif*, the total difference counter, by *edgeCounter.*

A loop causes the similarity and difference ratios to be retrieved from each index position of both arrays. *SumMean* is subtracted from the similarity ratio and stored in the local variable *v1*. *SumDif* is subtracted from the difference ratio and stored in the local variable *v2.* Two variables, *sumDev* sums the deviation scores from *v1* and *difDev* sums the deviation scores from *v2*. *SumDevProd* sums deviation products in the form: *v1* * *v2*. *SumStdDev* sums the standard deviation as *v1* squared. *difDevProd* sums the standard deviation as *v2* squared.

After all of the ratios in the arrays have been analyzed, *sumStdDev* reduces to the square root of the result of dividing itself by *edgeCounter*. Likewise, *difStdDev* also reduces to the square root of the result of dividing itself by *edgeCounter.* The scoring matrix position *(i,j)* is assigned the result of dividing *sumDevProd* by the product of *edgeCounte* and *sumStdDev* and *difStdDev*.

Figure 8.26 shows pseudo-code for the product-moment calculation.

```
Pseudo-code for productMoment(int i, int j)
{   State * ptr = sptr=getStart(i)
    EdgeNode *eptr = getFirstEdge(j)
    while( eptr )
    {    while ( eptr->isEdgeLeft() )
         {    temp = breakOffState()
              if(match)
              {    x= isChild(sptr, temp)
                   if(x)
                   {    sptr= getChild(x)
                        same++ }
                   else
                   {    match = false
                        different++     }    }
              else
                   different++ }
           edges = same + different
           sArray[edgeCounter]= same/edges
           dArray[edgeCounter]= different/edges
           sumEdges += edges
           sumSim += sArray[edgeCounter]
           sumDif += dArray[edgeCounter++]
           same = different = edges =  0
           sptr = corpus[i]->getStart()
           match = true
           eptr=eptr->next     }
      sumMean = sumSim/edgeCounter;
      difMean = sumDif/edgeCounter;
      for(c = 0; c < edgeCounter; c++)
      {    v1 = sArray[c]-sumMean
           v2 = dArray[c]-difMean
           sumDev += v1
           difDev += v2
           sumDevProd = sumDevProd + v1*v2
           sumStdDev = sumStdDev + ( v1*v1)
           difStdDev = difStdDev + ( v2*v2)    }
      sumStdDev = sqrt(sumStdDev/edgeCounter)
      difStdDev = sqrt(difStdDev/edgeCounter)
      if(sumEdges != 0)
           scoringMatrix[i][j]= sumDevProd/ edgeCounter * sumStdDev * difStdDev
}
```

Figure 8.26 Pseudo-Code for ProductMoment().

Each position in the scoring matrix holds the result of the equation shown in Figure 8.26, note that SM = *scoringMatrix*.

$$\text{SM[i][j]} = \frac{\text{sumDevProd}}{\text{NumPaths} * \text{simStdDev} * \text{difStdDev}}$$

Figure 8.26 Scoring Equation for Product Moment Calculation.

CHAPTER 9

EVALUATION

The five Reuters-21578 data sets, described in chapter 3, have been processed in groups of ten and fifty. The document sets evaluated in this chapter, reflect the processing of the first ten documents per data set. The document identification numbers used to populate the data sets are included as Appendix A-E. Result matrices for the fifty-document data sets are included as Appendix I-M.

For all of the result sets, larger numbers indicate greater similarity. Each row reflects how similar a document is to every other document. The values are asymmetrical and directional toward the documents with more saliency.

9.1 Contrast Model Results

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.22	-0.15	-0.14	-0.14	-0.14	0.18	-0.05	-0.05	-0.10
FSM 2	-0.38	1.00	-0.46	-0.28	-0.29	-0.27	-0.27	-0.27	-0.15	-0.45
FSM 3	-0.21	-0.37	1.00	-0.27	-0.20	-0.21	-0.13	-0.14	-0.17	-0.23
FSM 4	-0.27	-0.25	-0.33	1.00	-0.16	-0.23	-0.20	-0.13	-0.14	-0.30
FSM 5	-0.30	-0.33	-0.32	-0.22	1.00	0.03	-0.07	0.08	-0.22	-0.31
FSM 6	-0.38	-0.40	-0.43	-0.36	-0.08	1.00	-0.21	0.10	-0.33	-0.43
FSM 7	0.07	-0.21	-0.17	-0.14	0.07	0.05	1.00	0.22	0.06	-0.11
FSM 8	-0.27	-0.35	-0.31	-0.22	0.03	0.20	0.01	1.00	-0.24	-0.31
FSM 9	-0.19	-0.04	-0.24	-0.10	-0.13	-0.16	-0.01	-0.13	1.00	-0.28
FSM 10	-0.02	-0.25	-0.07	-0.03	-0.01	-0.05	0.08	0.05	-0.08	1.00

Figure 9.1 Data Set 1 – Contrast Model – Result Matrix.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	476.00	65.92	146.83	62.57	80.27	60.68	259.64	94.84	125.06	300.31
FSM 2	188.74	118.00	96.91	50.56	66.40	50.75	168.71	73.52	101.82	193.39
FSM 3	239.86	54.34	261.00	53.28	74.71	56.04	203.70	85.91	109.97	259.19
FSM 4	217.33	59.81	116.61	106.00	75.66	52.63	183.23	82.63	107.59	228.33
FSM 5	209.90	52.98	114.05	52.81	135.00	64.00	195.77	95.32	97.55	221.95
FSM 6	180.28	46.14	94.26	44.28	74.90	97.00	164.92	90.51	83.43	184.52
FSM 7	296.43	63.59	138.28	60.50	93.46	69.41	358.00	110.93	131.28	286.78
FSM 8	208.93	50.66	111.95	52.72	83.66	71.25	202.89	143.00	94.92	217.24
FSM 9	235.02	74.99	127.39	61.83	78.28	57.52	212.48	84.02	197.00	239.20
FSM 10	301.94	65.05	165.23	72.26	94.76	68.59	253.95	107.06	126.69	548.00

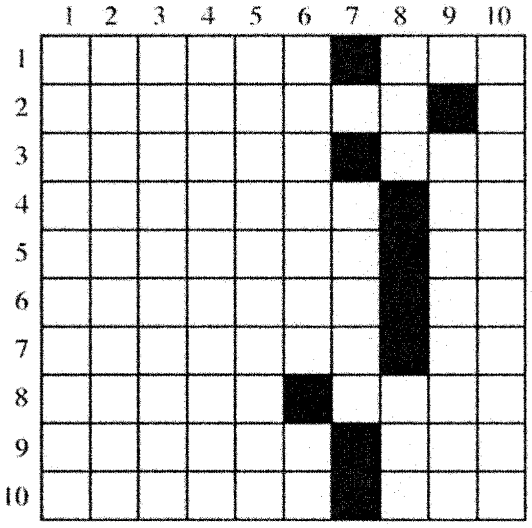
Figure 9.2 Data Set 1 – Contrast Model – Intersection Matrix.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	0.00	0.44	0.44	0.41	0.41	0.37	0.27	0.34	0.37	0.45
FSM 2	0.60	0.00	0.63	0.52	0.51	0.48	0.53	0.49	0.48	0.65
FSM 3	0.50	0.54	0.00	0.50	0.45	0.42	0.43	0.40	0.44	0.53
FSM 4	0.54	0.49	0.55	0.00	0.44	0.46	0.49	0.42	0.45	0.58
FSM 5	0.56	0.55	0.56	0.50	0.00	0.34	0.45	0.33	0.50	0.59
FSM 6	0.62	0.61	0.64	0.58	0.45	0.00	0.54	0.37	0.58	0.66
FSM 7	0.38	0.46	0.47	0.43	0.31	0.28	0.00	0.22	0.33	0.48
FSM 8	0.56	0.57	0.57	0.50	0.38	0.27	0.43	0.00	0.52	0.60
FSM 9	0.51	0.36	0.51	0.42	0.42	0.41	0.41	0.41	0.00	0.56
FSM 10	0.37	0.45	0.37	0.32	0.30	0.29	0.29	0.25	0.36	0.00

Figure 9.3 Data Set 1 – Contrast Model – Difference Matrix.

The contrast model is the simplest method of evaluation. It shows the semantic difference as the proportion of shared edges less the proportion of distinct edges. Figure 9.1 demonstrates the results from the first ten documents in Data Set 1. Figure 9.2 shows the values in the intersection matrix and Figure 9.3 shows the values in the difference matrix. The results in Figure 9.1 reflect the difference between the intersection matrix and the difference matrix. This measurement reflects the proportion of similarity between documents based on raw semantic distance.

Each row in the result matrix lists the similarity of one document to every other document. Following the assessment of all scores in the result matrix, the highest score for each document is marked in a similarity grid. Figure 9.4 shows the highest-ranking score based on the Contrast Model algorithm for documents in data set one. The highest-ranking score considers comparisons other than a document to itself.



	1	2	3	4	5	6	7	8	9	10
1							■			
2									■	
3							■			
4								■		
5								■		
6								■		
7								■		
8						■				
9							■			
10							■			

Figure 9.4 Data Set 1 – Contrast Model – Similarity Grid.

From the similarity grid, clusters of similar documents are derived by grouping documents vertically. Figure 9.5 shows the clusters created from the grid in 9.4. These clusters represent the final result for similarity assessment. Documents are identified by their data set identification number.

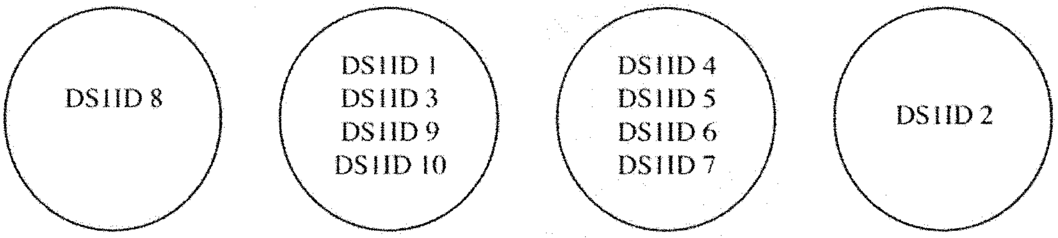


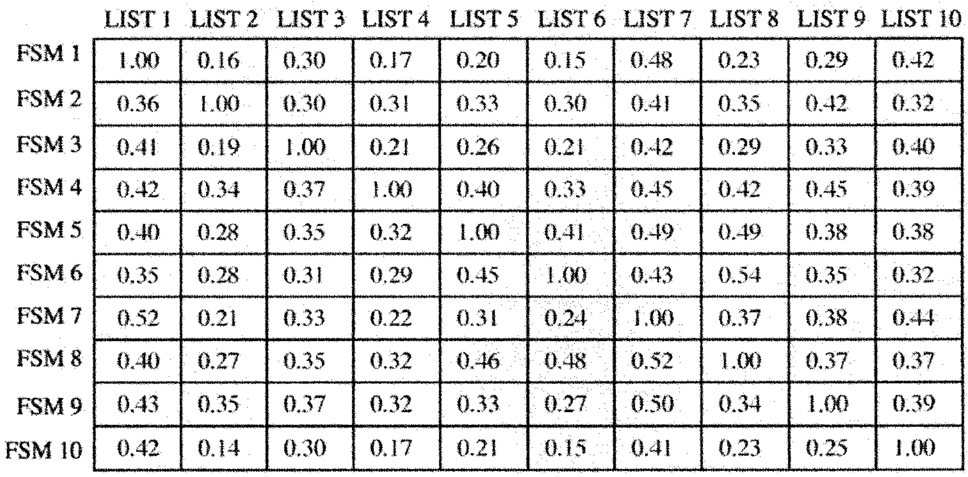
Figure 9.5 Data Set 1 – Contrast Model – Clusters of Similar Documents.

The result matrices, similarity grids, and similar document clusters for the remaining four data sets can be found in Appendix F.

9.2 Ratio Model Results

The results from the Ratio Model are normalized to fall between zero and one. The algorithm assesses the intersection of two sets by summing the proportion of matching edges to non-matching edges. The ratio of matching path edges to total edges is computed, as is the ratio of non-matching path edges to total edges. These values are figured for every document in the data set, and stored in a matching matrix and a difference matrix.

The Result matrix in Figure 9.6 shows the result of comparing each document a to document b as the matching ratio (*a*, *b*) divided by the sum of the matching ratio (*a*, *b*) added to the difference ratio (*a*, *b*) and the difference ratio (*b*, *a*). The result shows the percentage of similarity of one document to another, given the total difference.



	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.16	0.30	0.17	0.20	0.15	0.48	0.23	0.29	0.42
FSM 2	0.36	1.00	0.30	0.31	0.33	0.30	0.41	0.35	0.42	0.32
FSM 3	0.41	0.19	1.00	0.21	0.26	0.21	0.42	0.29	0.33	0.40
FSM 4	0.42	0.34	0.37	1.00	0.40	0.33	0.45	0.42	0.45	0.39
FSM 5	0.40	0.28	0.35	0.32	1.00	0.41	0.49	0.49	0.38	0.38
FSM 6	0.35	0.28	0.31	0.29	0.45	1.00	0.43	0.54	0.35	0.32
FSM 7	0.52	0.21	0.33	0.22	0.31	0.24	1.00	0.37	0.38	0.44
FSM 8	0.40	0.27	0.35	0.32	0.46	0.48	0.52	1.00	0.37	0.37
FSM 9	0.43	0.35	0.37	0.32	0.33	0.27	0.50	0.34	1.00	0.39
FSM 10	0.42	0.14	0.30	0.17	0.21	0.15	0.41	0.23	0.25	1.00

Figure 9.6 Data Set 1 – Ratio Model – Result Matrix.

Once the result matrix is complete, a similarity grid is generated by recording the highest-ratio scores. The grid for the first data set compared using the Ratio Model algorithm is shown in Figure 9.7. The last step in similarity assessment is completed by vertically grouping documents into clusters. The clusters formed by the similarity matrix in 9.7 shown in Figure 9.8.

	1	2	3	4	5	6	7	8	9	10
1							■			
2									■	
3							■			
4							■			
5								■		
6								■		
7	■									
8							■			
9							■			
10	■									

Figure 9.7 Data Set 1 – Ratio Model – Similarity Grid.

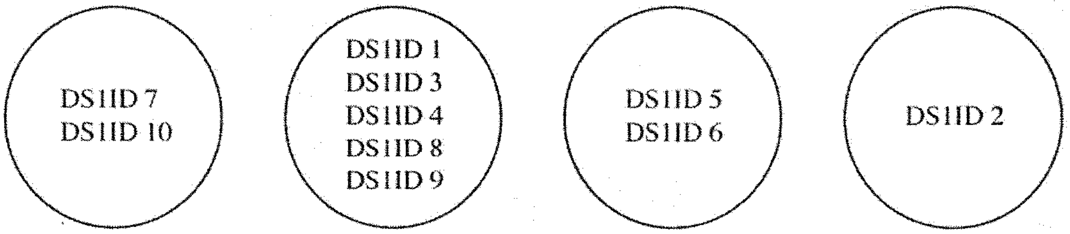


Figure 9.8 Data Set 1 – Ratio Model – Clusters of Similar Documents.

The result matrices, similarity grids, and similar document clusters for the remaining four data sets can be found in Appendix G.

9.3 Pearson Product Moment Coefficient Correlation

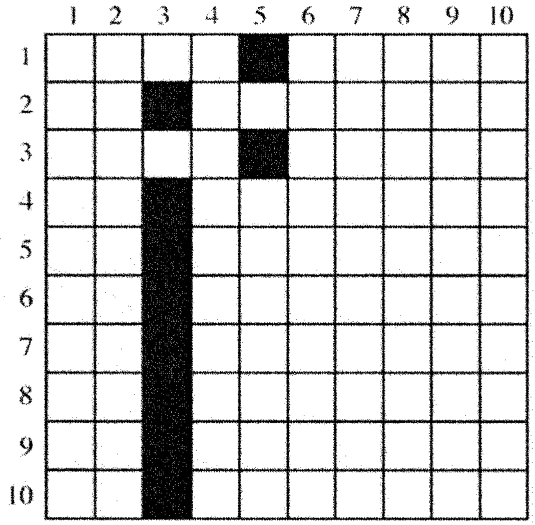
The product moment coefficient correlation is the most complex of all of the algorithms. The values reflect the degree to which two document's path edges match. In this model, the closer values come to zero–the smaller the correlation. A one indicates total correlation and a negative one indicates total negative correlation.

Figure 9.9 shows the result matrix generated for the first set of documents. Once all of the correlation scores are recorded, the results are mined for top correlation and a similarity grid is created. Figure 9.10 shows the grid created from the result matrix shown in Figure 9.9. The last step is to vertically group the similar documents into clusters. These clusters are shown in Figure 9.11.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.05	0.15	-0.31	0.18	0.16	0.01	-0.10	0.06	-0.08
FSM 2	0.13	1.00	0.15	-0.47	-0.02	-0.16	-0.05	-0.15	-0.16	0.09
FSM 3	0.13	-0.13	1.00	-0.23	0.18	0.01	0.10	-0.02	0.01	0.05
FSM 4	0.10	-0.21	0.27	1.00	-0.04	-0.12	0.02	-0.14	-0.17	0.11
FSM 5	0.15	-0.23	0.22	-0.49	1.00	-0.40	-0.20	-0.34	-0.20	-0.10
FSM 6	0.07	-0.27	0.09	-0.40	-0.35	1.00	-0.26	-0.45	-0.07	-0.06
FSM 7	-0.05	-0.36	0.13	-0.49	-0.27	-0.39	1.00	-0.43	-0.22	-0.26
FSM 8	-0.10	-0.33	0.06	-0.54	-0.34	-0.41	-0.36	1.00	-0.15	-0.22
FSM 9	0.08	-0.19	0.13	-0.55	-0.14	-0.08	-0.09	-0.08	1.00	0.01
FSM 10	-0.01	-0.15	0.09	-0.34	-0.30	-0.27	-0.20	-0.29	-0.10	1.00

Figure 9.9 Data Set 1 – Product Moment Model – Result Matrix.

The result matrices, similarity grids, and similar document clusters for the remaining four data sets can be found in Appendix H.



	1	2	3	4	5	6	7	8	9	10
1					■					
2			■							
3					■					
4			■							
5			■							
6			■							
7			■							
8			■							
9			■							
10			■							

Figure 9.10 Data Set 1 – Product Moment – Similarity Grid.

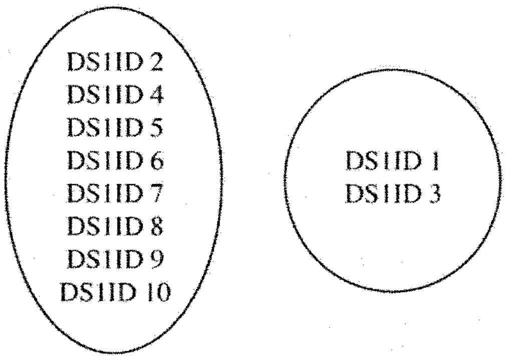


Figure 9.11 Data Set 1 – Product Moment – Clusters of Similar Documents.

9.4 What Does It Mean?

The numbers generated from each type of comparison describe a different method of scoring the similarity between documents. The Contrast Model relates the semantic distance between documents–calculated based on the difference between shared and distinct edges; the Ratio Model shows the proportion of matching edges between two documents; and the Product Moment Correlation reveals documents that are statistically similar–in shared and distinct edge distribution. For verification of accuracy, a closer look is provided.

9.4.1 Under the Magnifying Glass

The matrices of numbers, while impressive, aren't very telling of the process that generates them. A small-scale demonstration of the process that generates these numbers is presented in this section. This demonstration is designed to expose how documents are evaluated, by revealing in great detail how each document is processed. In this demonstration, a data set of five mini documents is compared. These documents consist of the first few sentences of the original Reuters documents. They are small, to keep the demonstration simple.

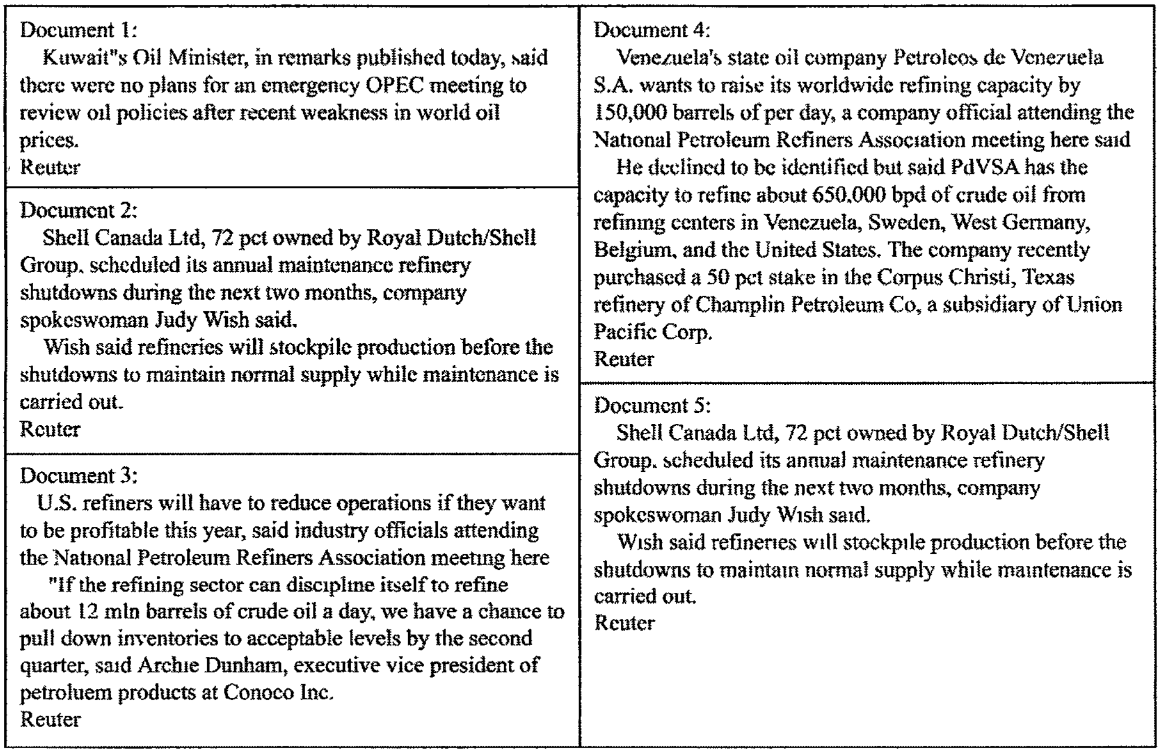
The second and fifth documents are identical. These documents match to test that each algorithm can identify exact matches. Topically, all of the documents are concerned with oil, although the specific term 'oil' does not appear in each document. The five documents used in this demonstration are shown in Figure 9.12.

The first step in processing breaks each document into word lists for semantic analysis. Figures 9.13-9.16 demonstrate the word list generated for each document and the statistics created from the semantic analysis of each document. (Note that document 2 is the same document as document 5, and is not represented twice.)

The 'Total Word Senses' heading in the Document Statistics section, indicates the total number of senses found through WordNet queries for all of the words in the document. This number reflects the number of paths that define the document. 'Total Number Edges' describes the total number of path pieces in a document.

'Analysis Words' indicates the number of words deemed fit for semantic analysis in the document. These are the words generated by the string trie traversal that are not proper nouns and are not stoplist words. The 'Words Analyzed' count shows the number of words out of the total 'Analysis Words' that are represented in the semantic network.

Words that fail to produce edge lists are included in the 'Semantics Analysis Statistics' with a brief description what went wrong. The most common reasons for paths not being added are words not located in WordNet, paths with cycles, and "instance-of" tags. Cycles are ignored because the edge list oscillates infinitely. At leaf levels, "instance-of" concepts are sporadically returned. These concepts are ignored because they are overly specific proper noun instances that lack conceptual relevance.



Document 1:

Kuwait"s Oil Minister, in remarks published today, said there were no plans for an emergency OPEC meeting to review oil policies after recent weakness in world oil prices.

Reuter

Document 2:

Shell Canada Ltd, 72 pct owned by Royal Dutch/Shell Group, scheduled its annual maintenance refinery shutdowns during the next two months, company spokeswoman Judy Wish said.

Wish said refineries will stockpile production before the shutdowns to maintain normal supply while maintenance is carried out.

Reuter

Document 3:

U.S. refiners will have to reduce operations if they want to be profitable this year, said industry officials attending the National Petroleum Refiners Association meeting here

"If the refining sector can discipline itself to refine about 12 mln barrels of crude oil a day, we have a chance to pull down inventories to acceptable levels by the second quarter, said Archie Dunham, executive vice president of petroluem products at Conoco Inc.

Reuter

Document 4:

Venezuela's state oil company Petroleos de Venezuela S.A. wants to raise its worldwide refining capacity by 150,000 barrels of per day, a company official attending the National Petroleum Refiners Association meeting here said

He declined to be identified but said PdVSA has the capacity to refine about 650,000 bpd of crude oil from refining centers in Venezuela, Sweden, West Germany, Belgium, and the United States. The company recently purchased a 50 pct stake in the Corpus Christi, Texas refinery of Champlin Petroleum Co, a subsidiary of Union Pacific Corp.

Reuter

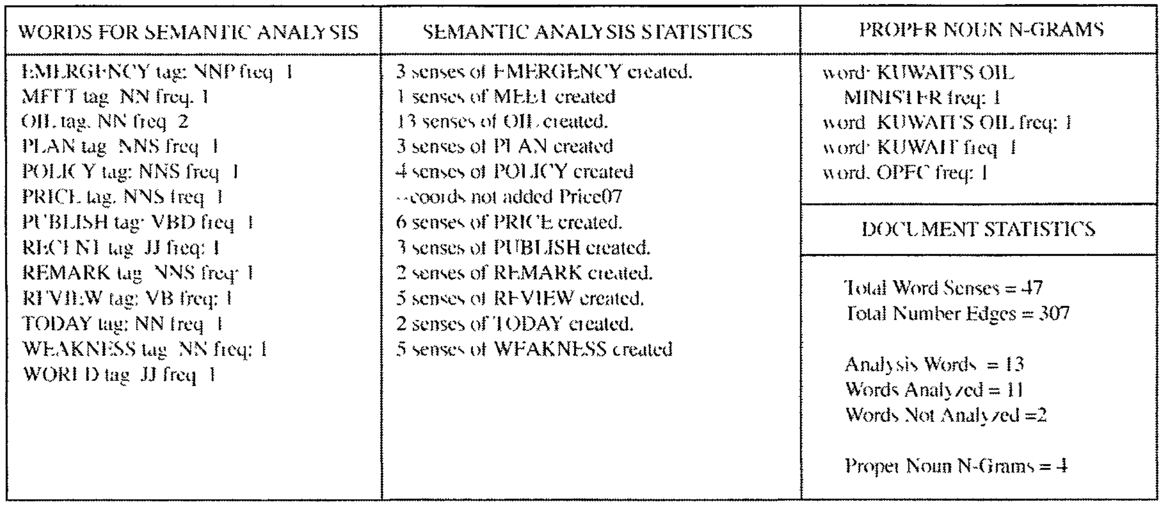
Document 5:

Shell Canada Ltd, 72 pct owned by Royal Dutch/Shell Group, scheduled its annual maintenance refinery shutdowns during the next two months, company spokeswoman Judy Wish said.

Wish said refineries will stockpile production before the shutdowns to maintain normal supply while maintenance is carried out.

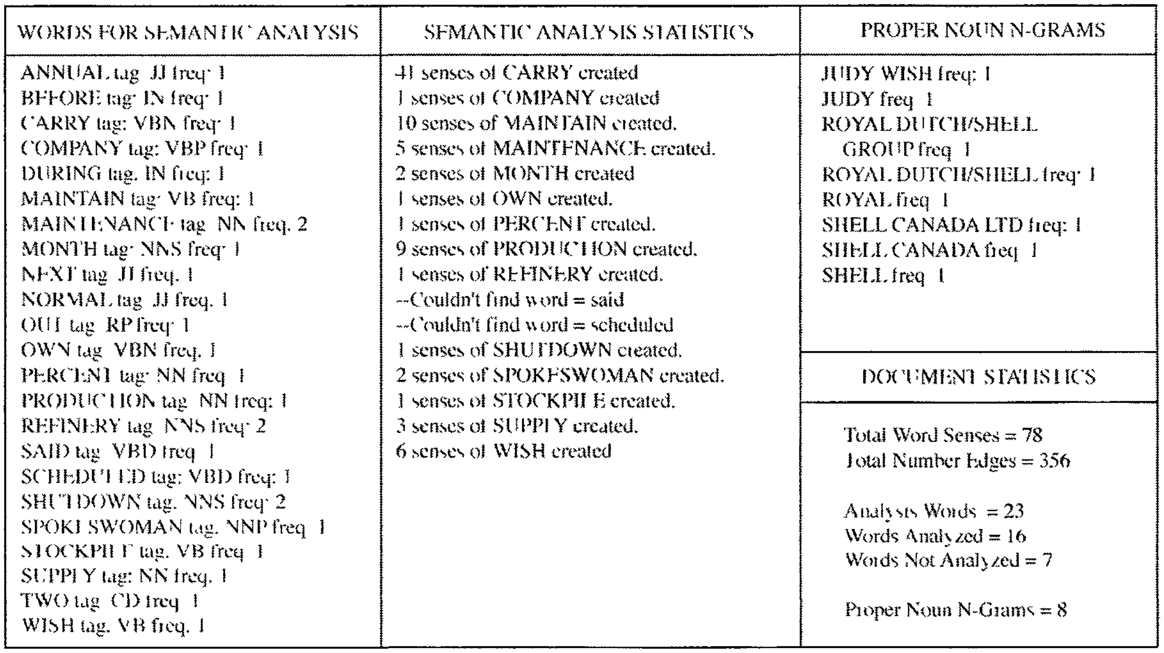
Reuter

Figure 9.12 Evaluation Data Set.



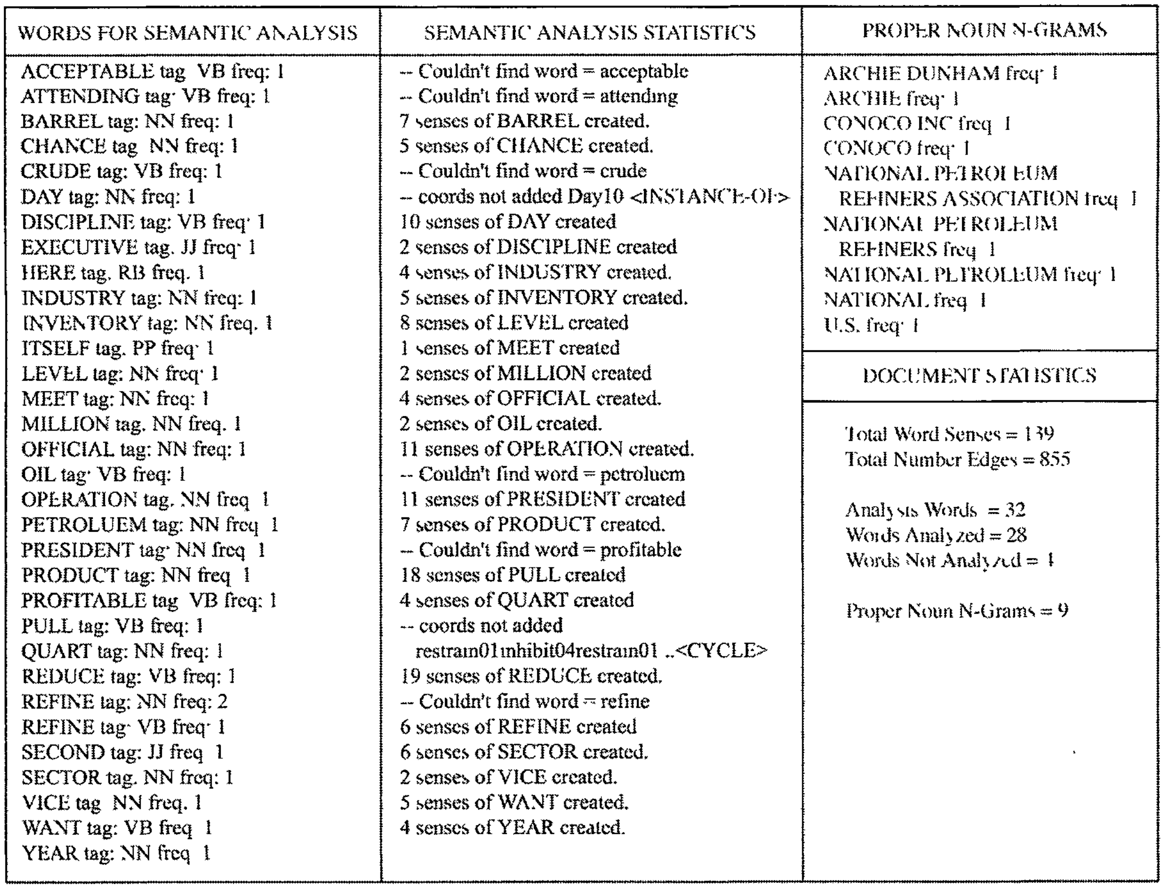
WORDS FOR SEMANTIC ANALYSIS	SEMANTIC ANALYSIS STATISTICS	PROPER NOUN N-GRAMS
EMERGENCY tag: NNP freq 1 MEET tag NN freq. 1 OIL tag. NN freq 2 PLAN tag NNS freq 1 POLICY tag: NNS freq 1 PRICE tag. NNS freq 1 PUBLISH tag: VBD freq 1 RECENT tag JJ freq: 1 REMARK tag NNS freq: 1 REVIEW tag: VB freq: 1 TODAY tag: NN freq 1 WEAKNESS tag NN freq: 1 WORLD tag JJ freq 1	3 senses of EMERGENCY created. 1 senses of MEET created 13 senses of OIL created. 3 senses of PLAN created 4 senses of POLICY created --coords not added Price07 6 senses of PRICE created. 3 senses of PUBLISH created. 2 senses of REMARK created. 5 senses of REVIEW created. 2 senses of TODAY created. 5 senses of WEAKNESS created	word: KUWAIT'S OIL MINISTER freq: 1 word KUWAIT'S OIL freq: 1 word: KUWAIT freq 1 word. OPEC freq: 1 DOCUMENT STATISTICS Total Word Senses = 47 Total Number Edges = 307 Analysis Words = 13 Words Analyzed = 11 Words Not Analyzed =2 Proper Noun N-Grams = 4

Figure 9.13 Analysis of Document 1.



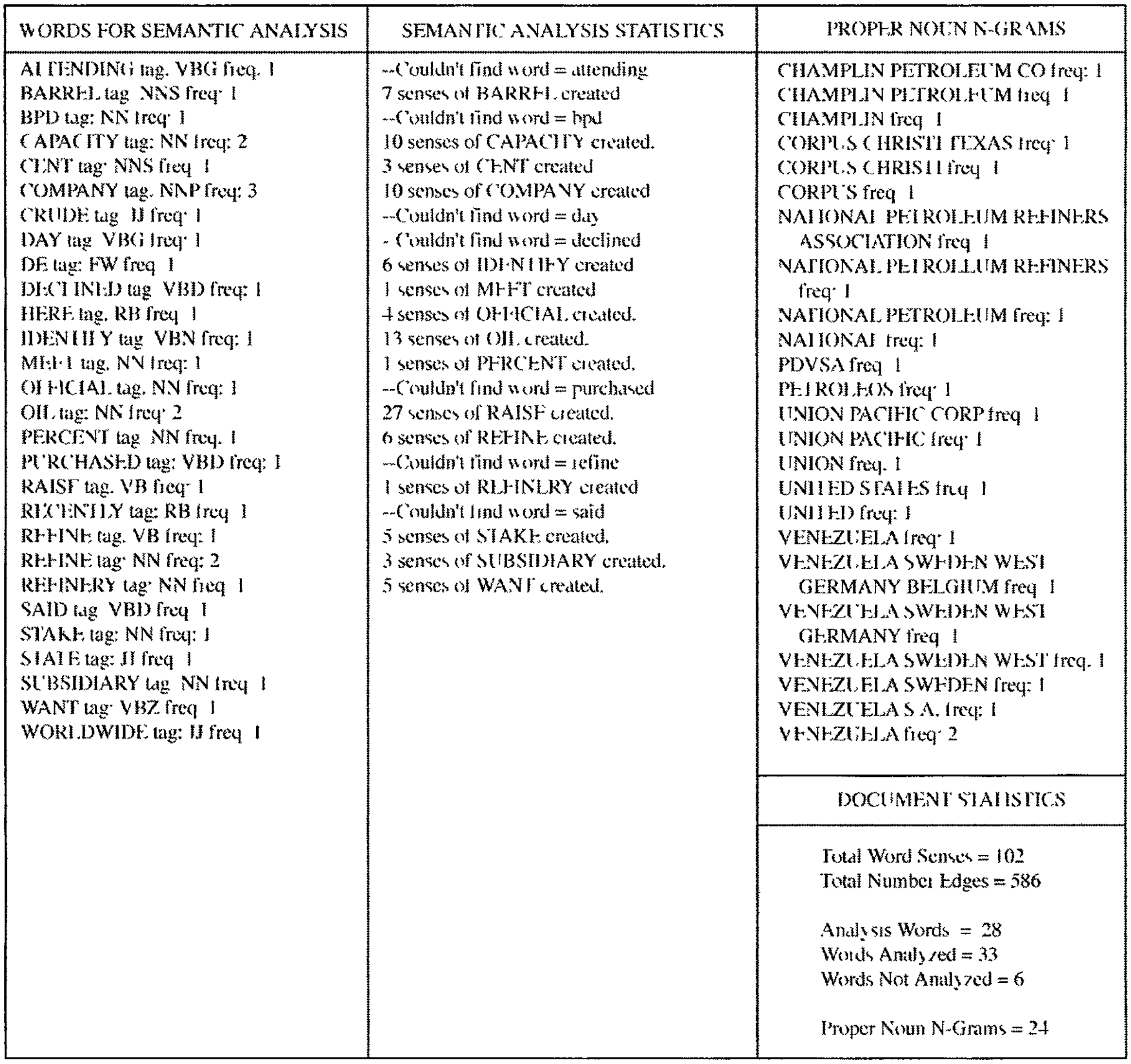
WORDS FOR SEMANTIC ANALYSIS	SEMANTIC ANALYSIS STATISTICS	PROPER NOUN N-GRAMS
ANNUAL tag JJ freq: 1 BEFORE tag: IN freq: 1 CARRY tag: VBN freq: 1 COMPANY tag: VBP freq: 1 DURING tag. IN freq: 1 MAINTAIN tag: VB freq: 1 MAINTENANCE tag NN freq. 2 MONTH tag: NNS freq: 1 NEXT tag JJ freq. 1 NORMAL tag JJ freq. 1 OUT tag RP freq: 1 OWN tag VBN freq. 1 PERCENT tag: NN freq 1 PRODUCTION tag NN freq: 1 REFINERY tag NNS freq: 2 SAID tag VBD freq 1 SCHEDULED tag: VBD freq: 1 SHUTDOWN tag. NNS freq: 2 SPOKESWOMAN tag. NNP freq 1 STOCKPILE tag. VB freq 1 SUPPLY tag: NN freq. 1 TWO tag CD freq 1 WISH tag. VB freq. 1	41 senses of CARRY created 1 senses of COMPANY created 10 senses of MAINTAIN created. 5 senses of MAINTENANCE created. 2 senses of MONTH created 1 senses of OWN created. 1 senses of PERCENT created. 9 senses of PRODUCTION created. 1 senses of REFINERY created. --Couldn't find word = said --Couldn't find word = scheduled 1 senses of SHUTDOWN created. 2 senses of SPOKESWOMAN created. 1 senses of STOCKPILE created. 3 senses of SUPPLY created. 6 senses of WISH created	JUDY WISH freq: 1 JUDY freq 1 ROYAL DUTCH/SHELL GROUP freq 1 ROYAL DUTCH/SHELL freq: 1 ROYAL freq 1 SHELL CANADA LTD freq: 1 SHELL CANADA freq 1 SHELL freq 1 DOCUMENT STATISTICS Total Word Senses = 78 Total Number Edges = 356 Analysis Words = 23 Words Analyzed = 16 Words Not Analyzed = 7 Proper Noun N-Grams = 8

Figure 9.14 Analysis of Document 2.



WORDS FOR SEMANTIC ANALYSIS	SEMANTIC ANALYSIS STATISTICS	PROPER NOUN N-GRAMS
ACCEPTABLE tag VB freq: 1	-- Couldn't find word = acceptable	ARCHIE DUNHAM freq· 1
ATTENDING tag· VB freq: 1	-- Couldn't find word = attending	ARCHIE freq· 1
BARREL tag: NN freq: 1	7 senses of BARREL created.	CONOCO INC freq 1
CHANCE tag NN freq: 1	5 senses of CHANCE created.	CONOCO freq· 1
CRUDE tag: VB freq: 1	-- Couldn't find word = crude	NATIONAL PETROLEUM REFINERS ASSOCIATION freq 1
DAY tag: NN freq: 1	-- coords not added Day10 <INSTANCE-OF>	
DISCIPLINE tag: VB freq· 1	10 senses of DAY created	NATIONAL PETROLEUM REFINERS freq 1
EXECUTIVE tag. JJ freq· 1	2 senses of DISCIPLINE created	
HERE tag. RB freq. 1	4 senses of INDUSTRY created.	NATIONAL PETROLEUM freq· 1
INDUSTRY tag: NN freq: 1	5 senses of INVENTORY created.	NATIONAL freq 1
INVENTORY tag: NN freq. 1	8 senses of LEVEL created	U.S. freq· 1
ITSELF tag. PP freq· 1	1 senses of MEET created	
LEVEL tag: NN freq· 1	2 senses of MILLION created	DOCUMENT STATISTICS
MEET tag: NN freq: 1	4 senses of OFFICIAL created.	
MILLION tag. NN freq. 1	2 senses of OIL created.	Total Word Senses = 139
OFFICIAL tag: NN freq: 1	11 senses of OPERATION created.	Total Number Edges = 855
OIL tag· VB freq: 1	-- Couldn't find word = petroluem	
OPERATION tag. NN freq 1	11 senses of PRESIDENT created	Analysis Words = 32
PETROLUEM tag: NN freq 1	7 senses of PRODUCT created.	Words Analyzed = 28
PRESIDENT tag· NN freq 1	-- Couldn't find word = profitable	Words Not Analyzed = 4
PRODUCT tag: NN freq 1	18 senses of PULL created	
PROFITABLE tag VB freq: 1	4 senses of QUART created	Proper Noun N-Grams = 9
PULL tag: VB freq: 1	-- coords not added restrain01inhibit04restrain01 ..<CYCLE>	
QUART tag: NN freq: 1		
REDUCE tag: VB freq: 1	19 senses of REDUCE created.	
REFINE tag: NN freq: 2	-- Couldn't find word = refine	
REFINE tag· VB freq· 1	6 senses of REFINE created	
SECOND tag: JJ freq 1	6 senses of SECTOR created.	
SECTOR tag. NN freq: 1	2 senses of VICE created.	
VICE tag NN freq. 1	5 senses of WANT created.	
WANT tag: VB freq 1	4 senses of YEAR created.	
YEAR tag: NN freq 1		

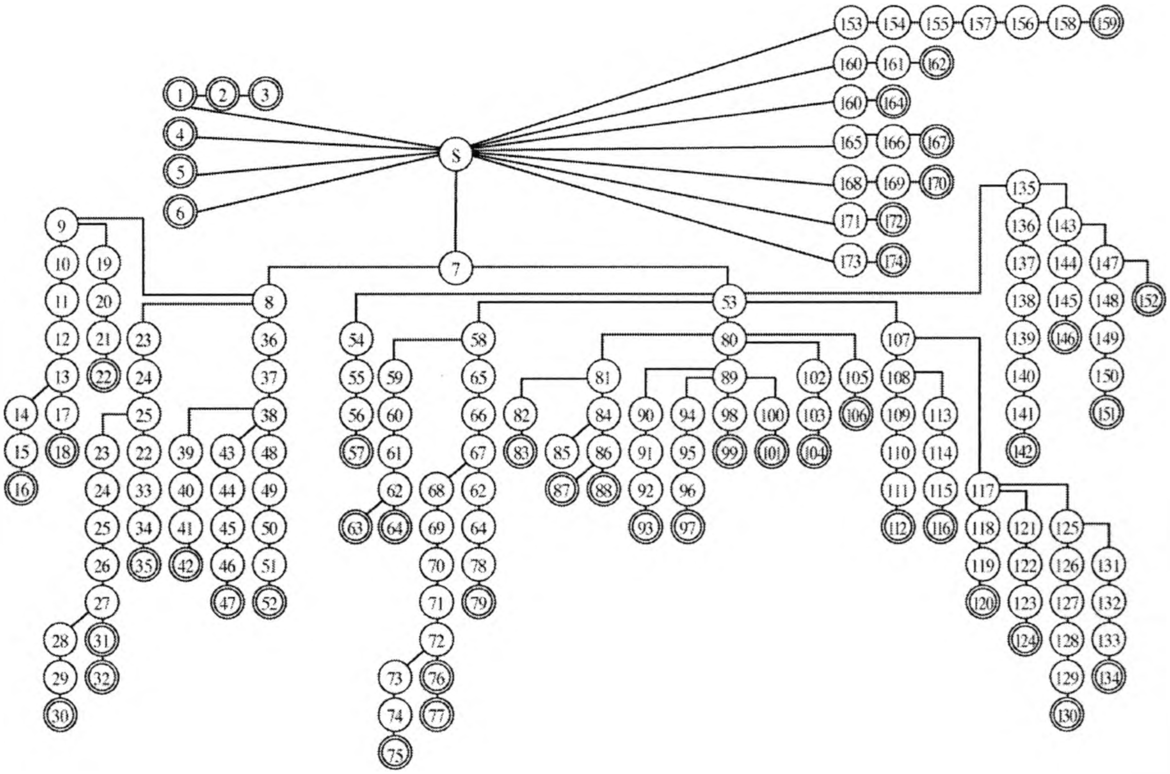
Figure 9.15 Analysis of Document 3.



WORDS FOR SEMANTIC ANALYSIS	SEMANTIC ANALYSIS STATISTICS	PROPER NOUN N-GRAMS
ATTENDING tag: VBG freq: 1	--Couldn't find word = attending	CHAMPLIN PETROLEUM CO freq: 1
BARREL tag: NNS freq: 1	7 senses of BARREL created	CHAMPLIN PETROLEUM freq: 1
BPD tag: NN freq: 1	--Couldn't find word = bpd	CHAMPLIN freq: 1
CAPACITY tag: NN freq: 2	10 senses of CAPACITY created.	CORPUS CHRISTI TEXAS freq: 1
CENT tag: NNS freq: 1	3 senses of CENT created	CORPUS CHRISTI freq: 1
COMPANY tag: NNP freq: 3	10 senses of COMPANY created	CORPUS freq: 1
CRUDE tag: JJ freq: 1	--Couldn't find word = day	NATIONAL PETROLEUM REFINERS ASSOCIATION freq: 1
DAY tag: VBG freq: 1	--Couldn't find word = declined	NATIONAL PETROLEUM REFINERS freq: 1
DE tag: FW freq: 1	6 senses of IDENTIFY created	NATIONAL PETROLEUM freq: 1
DECLINED tag: VBD freq: 1	1 senses of MEET created	NATIONAL freq: 1
HERE tag: RB freq: 1	4 senses of OFFICIAL created.	PDVSA freq: 1
IDENTIFY tag: VBN freq: 1	13 senses of OIL created.	PETROLEOS freq: 1
MEET tag: NN freq: 1	1 senses of PERCENT created.	UNION PACIFIC CORP freq: 1
OFFICIAL tag: NN freq: 1	--Couldn't find word = purchased	UNION PACIFIC freq: 1
OIL tag: NN freq: 2	27 senses of RAISE created.	UNION freq: 1
PERCENT tag: NN freq: 1	6 senses of REFINE created.	UNITED STATES freq: 1
PURCHASED tag: VBD freq: 1	--Couldn't find word = refine	UNITED freq: 1
RAISE tag: VB freq: 1	1 senses of REFINERY created	VENEZUELA freq: 1
RECENTLY tag: RB freq: 1	--Couldn't find word = said	VENEZUELA SWEDEN WEST GERMANY BELGIUM freq: 1
REFINE tag: VB freq: 1	5 senses of STAKE created.	VENEZUELA SWEDEN WEST GERMANY freq: 1
REFINE tag: NN freq: 2	3 senses of SUBSIDIARY created.	VENEZUELA SWEDEN WEST freq: 1
REFINERY tag: NN freq: 1	5 senses of WANT created.	VENEZUELA SWEDEN freq: 1
SAID tag: VBD freq: 1		VENEZUELA S A. freq: 1
STAKE tag: NN freq: 1		VENEZUELA freq: 2
STATE tag: JJ freq: 1		DOCUMENT STATISTICS
SUBSIDIARY tag: NN freq: 1		Total Word Senses = 102 Total Number Edges = 586
WANT tag: VBZ freq: 1		Analysis Words = 28 Words Analyzed = 33 Words Not Analyzed = 6
WORLDWIDE tag: JJ freq: 1		Proper Noun N-Grams = 24

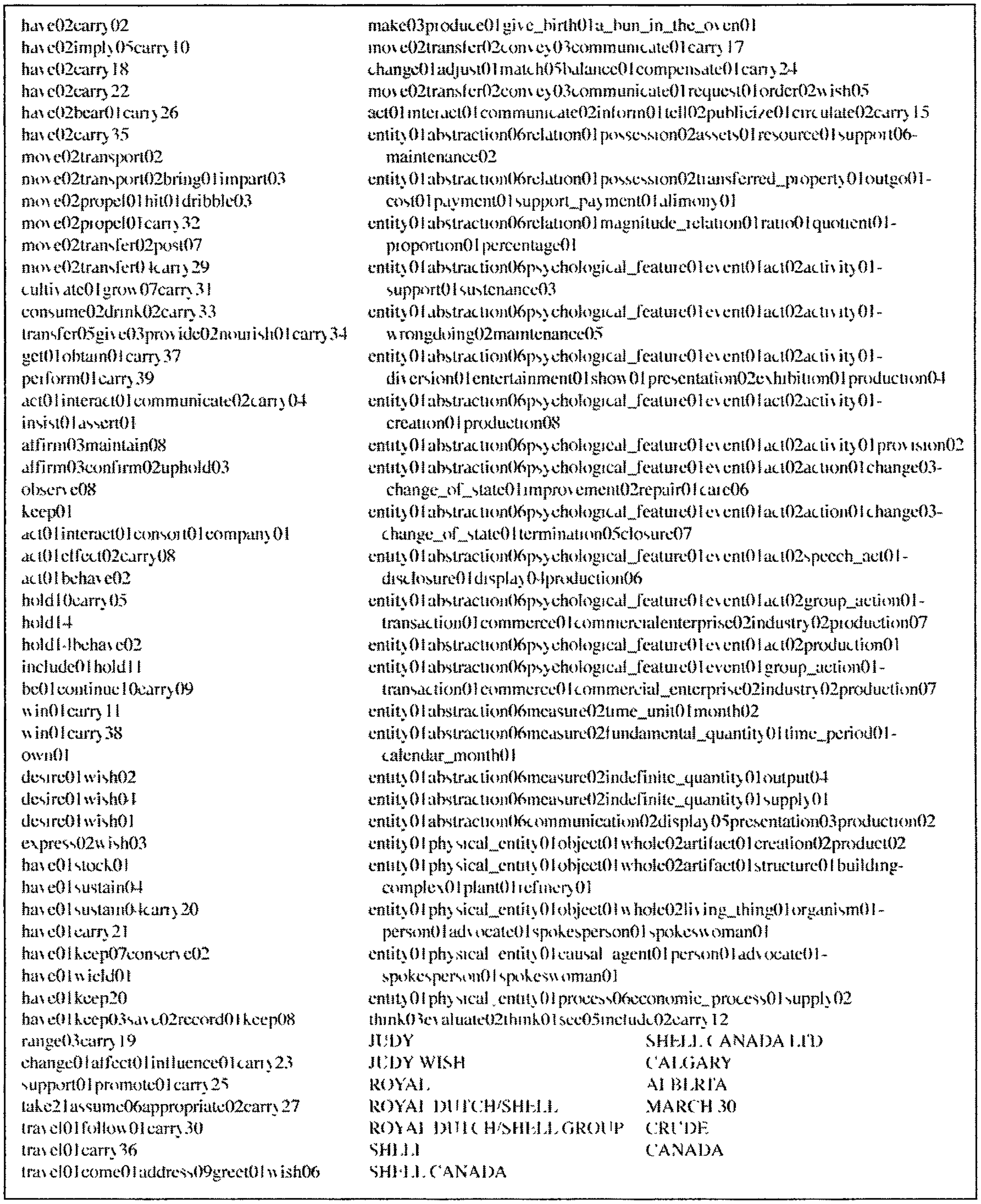
Figure 9.16 Analysis of Document 4.

Figure 9.17 illustrates a map of the finite state machine created for document one. Each edge piece becomes a state in the FSM–final edge pieces become accepting states. This FSM represents the semantic network for document one. Due to size constraints, each state is numbered. The state names that correspond to the numeric state labels are listed below the state diagram. 'S' indicates the *start* state.



1. KUWAIT'S
2. OIL
3. MINISTER
4. OPEC
5. ECUADOR
6. CRUDE
7. entity01
8. physical_entity01
9. thing12
10. unit05
11. molecule01
12. macromolecule01
13. lipid01
14. fat01
15. edible_fat01
16. vegetable_oil01
17. oil01
18. petroleum01
19. substance07
20. fuel01
21. fossil_fuel01
22. petroleum01
23. chemical01
24. compound02
25. organic_compound01
26. macromolecule01
27. lipid01
28. fat01
29. edible_fat01
30. vegetable_oil01
31. oil01
32. petroleum01
33. paint01
34. oil_paint01
35. oil02
36. object01
37. whole02
38. artifact01
39. creation02
40. representation02
41. drawing02
42. plan03
43. covering02
44. coating01
45. paint01
46. oil_paint01
47. oil02
48. instrumentality03
49. device01
50. restraint06
51. brake01
52. hand_brake01
53. abstraction06
54. measure02
55. time_unit01
56. day01
57. today02
58. relation01
59. possession02
60. transferred_property01
61. outgo01
62. cost01
63. price06
64. price02
65. part01
66. substance01
67. material01
68. chemical0
69. compound02
70. organic_compound011
71. macro-molecule01
72. lipid01
73. fat01
74. edible_fat01
75. vegetable_oil01
76. oil01
77. petroleum01
78. oil_paint01
79. oil02
80. attribute02
81. quality0
82. powerlessness01
83. helplessness01
84. worth02
85. price03
86. price04
87. monetary_value01
88. price03
89. state02
90. feeling01
91. liking01
92. preference01
93. weakness05
94. condition03
95. fortune04
96. misfortune02
97. weakness04
98. temporary_state01
99. emergency02
100. imperection01
101. failing01
102. time05
103. present01
104. today01
105. property02
106. weakness03
107. psychological_feature01
108. event01
109. happening01
110. juncture01
111. crisis02
112. emergency01
113. social_event01
114. contest01
115. athletic_contest01
116. meet01
117. cognition01
118. structure03
119. arrangement03
120. design02
121. content05
122. idea01
123. plan01
124. plan_of_action01
125. process02
126. higher_cog_process01
127. thinking01
128. reasoning01
129. argumentation02
130. policy02
131. basic_cog_process01
132. attention01
133. notice02
134. remark02
135. communication02
136. written_comm01
137. writing02
138. document01
139. legal_document01
140. written_agreement01
141. contract01
142. policy03
143. message02
144. offer02
145. reward04
146. price05
147. statement01
148. agreement01
149. written_agreement01
150. contract01
151. policy03
152. remark01
153. act01
154. interact01
155. communicate02
156. inform01
157. tell02
158. publicize01
159. publish02
160. make03
161. create_verbally01
162. publish02
163. produce02
164. print01
165. think03
166. evaluate02
167. review02
168. examine02
169. inspect01
170. review03
171. remember01
172. review04
173. analyze01
174. review01

Figure 9.17 Finite State Machine for Document One.



```
have02carry02
have02imply05carry10
have02carry18
have02carry22
have02bear01carry26
have02carry35
move02transport02
move02transport02bring01impart03
move02propel01hit01dribble03
move02propel01carry32
move02transfer02post07
move02transfer04carry29
cultivate01grow07carry31
consume02drink02carry33
transfer05give03provide02nourish01carry34
get01obtain01carry37
perform01carry39
act01interact01communicate02carry04
insist01assert01
affirm03maintain08
affirm03confirm02uphold03
observe08
keep01
act01interact01consort01company01
act01effect02carry08
act01behave02
hold10carry05
hold14
hold14behave02
include01hold11
be01continue10carry09
win01carry11
win01carry38
own01
desire01wish02
desire01wish04
desire01wish01
express02wish03
have01stock01
have01sustain04
have01sustain04carry20
have01carry21
have01keep07conserve02
have01wield01
have01keep20
have01keep03save02record01keep08
range03carry19
change01affect01influence01carry23
support01promote01carry25
take21assume06appropriate02carry27
travel01follow01carry30
travel01carry36
travel01come01address09greet01wish06
make03produce01give_birth01a_bun_in_the_oven01
move02transfer02convey03communicate01carry17
change01adjust01match05balance01compensate01carry24
move02transfer02convey03communicate01request01order02wish05
act01interact01communicate02inform01tell02publicize01circulate02carry15
entity01abstraction06relation01possession02assets01resource01support06-
   maintenance02
entity01abstraction06relation01possession02transferred_property01outgo01-
   cost01payment01support_payment01alimony01
entity01abstraction06relation01magnitude_relation01ratio01quotient01-
   proportion01percentage01
entity01abstraction06psychological_feature01event01act02activity01-
   support01sustenance03
entity01abstraction06psychological_feature01event01act02activity01-
   wrongdoing02maintenance05
entity01abstraction06psychological_feature01event01act02activity01-
   diversion01entertainment01show01presentation02exhibition01production04
entity01abstraction06psychological_feature01event01act02activity01-
   creation01production08
entity01abstraction06psychological_feature01event01act02activity01provision02
entity01abstraction06psychological_feature01event01act02action01change03-
   change_of_state01improvement02repair01care06
entity01abstraction06psychological_feature01event01act02action01change03-
  change_of_state01termination05closure07
entity01abstraction06psychological_feature01event01act02speech_act01-
  disclosure01display04production06
entity01abstraction06psychological_feature01event01act02group_action01-
  transaction01commerce01commercialenterprise02industry02production07
entity01abstraction06psychological_feature01event01act02production01
entity01abstraction06psychological_feature01event01group_action01-
  transaction01commerce01commercial_enterprise02industry02production07
entity01abstraction06measure02time_unit01month02
entity01abstraction06measure02fundamental_quantity01time_period01-
  calendar_month01
entity01abstraction06measure02indefinite_quantity01output04
entity01abstraction06measure02indefinite_quantity01supply01
entity01abstraction06communication02display05presentation03production02
entity01physical_entity01object01whole02artifact01creation02product02
entity01physical_entity01object01whole02artifact01structure01building-
  complex01plant01refinery01
entity01physical_entity01object01whole02living_thing01organism01-
  person01advocate01spokesperson01spokeswoman01
entity01physical_entity01causal_agent01person01advocate01-
  spokesperson01spokeswoman01
entity01physical_entity01process06economic_process01supply02
think03evaluate02think01see05include02carry12
JUDY
JUDY WISH
ROYAL
ROYAL DUTCH/SHELL
ROYAL DUTCH/SHELL GROUP
SHELL
SHELL CANADA
SHELL CANADA LTD
CALGARY
ALBERTA
MARCH 30
CRUDE
CANADA
```

Figure 9.18 Edge List Created from Document Two.

All three algorithms share the first step of comparing the FSM of one document to the edge list of another and counting the common and distinct edges. Figure 9.18 shows the elements of the edge list created for document two.

Figure 9.19 shows the FSM of document one with the common edges highlighted as grey states. Common edges have been found through the comparison of the edge list of the second document to the FSM of the first. The labels for this FSM correspond directly to the labels presented in Figure 9.17.

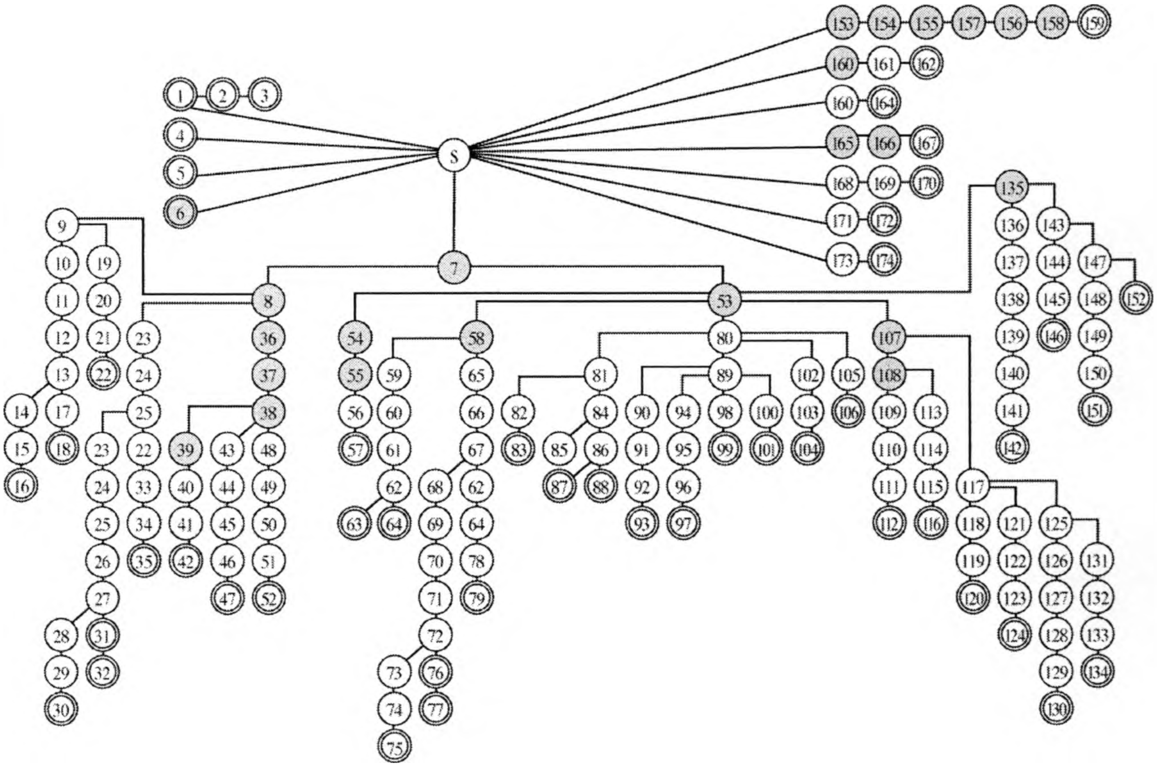


Figure 9.19 Finite State Machine Showing Edge Overlap.

Once all of the documents have been analyzed, they are evaluated for similarity based on one of the three algorithms: Contrast Model, Ratio Model, or Product Moment. Figure 9.20 shows the paths that generated the matches and the ratio of edge matches per matching path. These ratios become the input data to the three comparison models. From these ratios, document similarity is assessed.

Path Match Ratio	Matching Paths
3/4	act01interact01communicate02carry04
6/8	act01interact01communicate02inform01tell02publicize01circulate02carry15
2/4	act01interact01consort01company01
1/3	act01effect02carry08
1/2	act01behave02
2/6	think03evaluate02think01see05include02carry12
1/4	make03produce01give_birth01have_a_bun_in_the_oven01
4/8	entity01abstraction06relation01possession02assets01resource01support06maintenance02
7/10	entity01abstraction06relation01possession02transferred_property01outgo01cost01payment01support_payment01alimony01
3/5	entity01abstraction06relation01magnitude_relation01ratio01quotient01proportion01percentage01
4/8	entity01abstraction06psychological_feature01event01act02activity01support01sustenance03
4/8	entity01abstraction06psychological_feature01event01act02activity01wrongdoing02maintenance05
4/8	entity01abstraction06psychological_feature01event01act02activity01diversion01entertainment01show01presentation02exhibition01production04
4/8	entity01abstraction06psychological_feature01event01act02activity01creation01production08
4/7	entity01abstraction06psychological_feature01event01act02activity01provision02
4/7	entity01abstraction06psychological_feature01event01act02action01change03change_of_state01improvement02repair01care06
4/10	entity01abstraction06psychological_feature01event01act02action01change03change_of_state01termination05closure07
4/9	entity01abstraction06psychological_feature01event01act02speech_act01disclosure01display04production06
4/11	entity01abstraction06psychological_feature01event01act02group_action01transaction01commerce01commercial_enterprise02industry02production07
4/6	entity01abstraction06psychological_feature01event01act02production01
4/10	entity01abstraction06psychological_feature01event01group_action01transaction01commerce01commercial_enterprise02industry02production07
4/5	entity01abstraction06measure02time_unit01month02
3/6	entity01abstraction06measure02fundamental_quantity01time_period01calendar_month01
3/5	entity01abstraction06measure02indefinite_quantity01output04
3/5	entity01abstraction06measure02indefinite_quantity01supply01
3/6	entity01abstraction06communication02display05presentation03production02
6/7	entity01physical_entity01object01whole02artifact01creation02product02
5/9	entity01physical_entity01object01whole02artifact01structure01building_complex01plant01refinery01
4/10	entity01physical_entity01object01whole02living_thing01organism01person01advocate01spokesperson01spokeswoman01
2/7	entity01physical_entity01causal_agent01person01advocate01spokesperson01spokeswoman01
2/5	entity01physical_entity01process06economic_process01supply02
1/1	CRUDE

Figure 9.20 Ratios of Matching Path Prefixes.

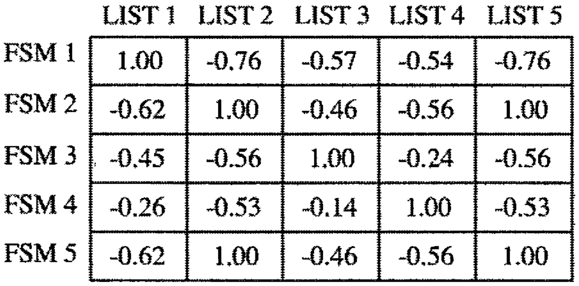
9.4.2 Contrast Model Comparison

The Contrast Model begins by calculating the sum of same edge to total edge ratios. These ratios are the matching prefixes of edge lists. The second document's same edge to total edge ratios are shown in the first column of Figure 9.28. In this case, the sum of the ratios is 17.03. This value is stored in the *similar* matrix.

Similarly, the document's different edge to total edge ratios are summed and then normalized by dividing by the total number of edge lists in the document. Continuing the example, the summed value for the non-matching path pieces is 78.97, this value is divided by the total number of edges in the document. Since there are 96 edges in the document, the normalized value, .82, is stored in the *difference* matrix.

The Contrast Model finds the intersection of two documents by multiplying the similar values of a document *a* to document *b* to the similar values of a document *b* to document *a*. This value is divided by the product of the total number of edges in document *a* times the total number of edges in document *b*. The value stored for the intersection of document *one* to document *two* is .06

The last computation in the Contrast Model subtracts from the intersection value of document *a* the difference value of document *b*. The value arrived at when comparing the edge list of document two to the FSM in document one is -.77. Figure 9.21 shows the Contrast Model result matrix for the five documents presented in 9.4. Figure 9.22 shows the similarity grid generated from these results and Figure 9.23 shows the clusters of similar documents, per the Contrast Model algorithm.



	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5
FSM 1	1.00	-0.76	-0.57	-0.54	-0.76
FSM 2	-0.62	1.00	-0.46	-0.56	1.00
FSM 3	-0.45	-0.56	1.00	-0.24	-0.56
FSM 4	-0.26	-0.53	-0.14	1.00	-0.53
FSM 5	-0.62	1.00	-0.46	-0.56	1.00

Figure 9.21 Contrast Model Result Matrix for Documents 1-5.

	1	2	3	4	5
1				■	
2			■		
3				■	
4			■		
5			■		

Figure 9.22 Documents 1-5 – Contrast Model – Similarity Grid.

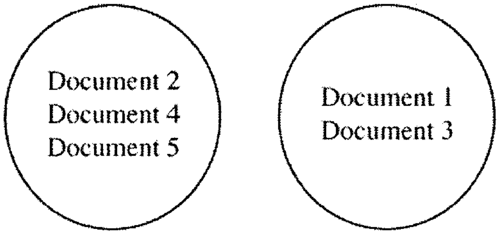


Figure 9.23 Documents 1-5 – Contrast Model – Clusters of Similar Documents.

9.4.3 Ratio Model Comparison

The Ratio Model begins much like the Contrast Model–it calculates the sum of same edge to total edge ratios. These ratios represent the matching prefixes of a document's edge list compared to a document's FSM. Document two's same-edge to total-edge ratios are shown in the first column of Figure 9.24. This value, 17.03 in this case, is stored in the *similar* matrix. Likewise, the different edge to total edge ratios are summed and stored in the *difference* matrix. The value stored in the *difference* matrix for this example, is 78.97.

The second phase of the Ration Model scores the similarity of an edge list in document *a* to the FSM in document *b*. The score is achieved by dividing the value in the intersection matrix for the ratio of shared edges between the edge list in document *a* and the FSM in document *b,* by the product of: the values stored in the intersection matrix for the shared edges between the edge list in document *b* and the FSM in document *a;* multiplied by the value stored in the difference matrix for different edges between the edge list in document *a* and the FSM in document *b;* multiplied by the value stored in the difference matrix for different edges between the edge list in document *b* and the FSM in document *a.*

The value stored for the comparison of the edge list of the second document compared to the FSM in the first is .13. Figure 9.24 shows the Ratio Model result matrix for the five documents presented in 9.4. Figure 9.25 shows the similarity grid generated from these results and Figure 9.26 shows the clusters of similar documents, per the Ratio Model algorithm.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5
FSM 1	1.00	0.13	0.25	0.25	0.13
FSM 2	0.13	1.00	0.29	0.22	1.00
FSM 3	0.14	0.16	1.00	0.32	0.16
FSM 4	0.21	0.18	0.40	1.00	0.18
FSM 5	0.13	1.00	0.29	0.22	1.00

Figure 9.24 Ratio Model Result Matrix for Documents 1-5.

	1	2	3	4	5
1			■		
2			■		
3				■	
4			■		
5			■		

Figure 9.25 Documents 1-5 – Ratio Model – Similarity Grid.

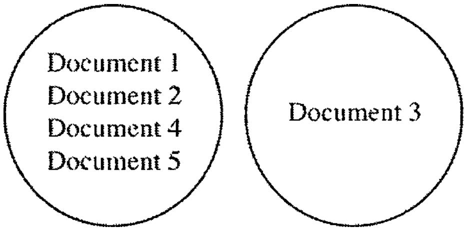


Figure 9.26 Documents 1-5 – Ratio Model – Clusters of Similar Documents.

9.4.4 Product Moment Comparison

The Product Moment comparison begins by comparing each path in on document's edge list to the FSM of another document. A count is maintained of the number of matching path edges and the number of non-matching path edges for each path. These counts are stored separately in two arrays.

After all of the edges in a document are counted, the arrays are processed. The mean is found for matching edges by summing the total number of matches and dividing by the total number of paths. In the example looking at the comparison of the edge list from the second document to the FSM of the first, the mean for matching edges is 1.16. The mean is found in the same way for non-matching edges, in this example the mean for non-matching edges is 2.86.

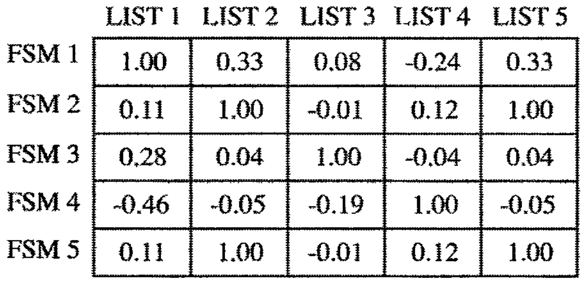
The value of the difference between each matching edge count and the matching mean is summed for all edges and stored in a local variable, *sumDev*. Similarly, the value for the difference between each non-matching edge count and the difference mean is summed for all edges and stored in a different local variable, *difDev*. The value of *sumDev* in this example results in 316.66 and the value of *difDev* results in 243.76.

The sum of each matching edge deviation score is squared and stored in the variable, *sumStdDev*. The sum of each non-matching edge deviation score is likewise squared and stored in a different variable, *difStdDev*. The sum of the product of each matching edge deviation and non-matching edge deviation is stored in yet another variable, *sumDevProd.*

Once all of the deviation scores have been calculated, the standard deviation is found for the matching edges and non-matching edges. The standard deviation value for matching edges is arrived at by taking the square root of the value produced by dividing *sumStdDev* by the number of edges in the document. The standard deviation for *difStdDev* is found similarly. In this example, *sumStdDev* is 1.82 and *difStdDev* is 1.60.

The product moment score is the result of dividing *sumDevProd* by the product of the number of edge lists in the document multiplied by *sumStdDev* multiplied by *difStdDev*. In this example, the product moment score expressing the correlation between the edge list of the second document with the FSM of the first is .33.

Figures 9.27, 9.28, and 9.29 show the Product Moment result matrix, similarity grid and document clusters created for the five documents presented in 9.4.



	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5
FSM 1	1.00	0.33	0.08	-0.24	0.33
FSM 2	0.11	1.00	-0.01	0.12	1.00
FSM 3	0.28	0.04	1.00	-0.04	0.04
FSM 4	-0.46	-0.05	-0.19	1.00	-0.05
FSM 5	0.11	1.00	-0.01	0.12	1.00

Figure 9.27 Product Moment Result Matrix for Documents 1-5.

	1	2	3	4	5
1	□	■	□	□	■
2	□	□	□	■	□
3	■	□	□	□	□
4	□	■	□	□	■
5	□	□	□	■	□

Figure 9.28 Documents 1-5 – Product Moment – Similarity Grid.

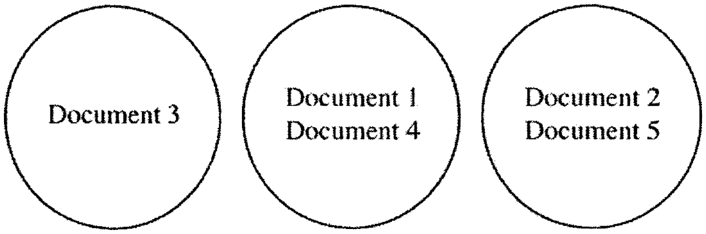


Figure 9.29 Documents 1-5 – Product Moment – Clusters of Similar Documents.

9.5 Modifications to EdgeNode List

The initial EdgeNode list contained all of the edges found from semantic analysis through WordNet. The original thought was that multiple hits would increase the weight of an edge. But this thought was tempered by a study referenced by Robert Nosofsky. This study questioned under what conditions frequency was inversely related to similarity. It showed that frequent exposure to less-salient members of a category performed worse in recognition, than less-frequent exposure to highly salient members [10, 54].

To increase the salience of the features and reduce the noise, the EdgeNode list has been reduced. The FSM is used to build a new EdgeNode list, which contains a minimal feature. This more concise EdgeNode list is used in similarity comparisons.

9.6 Modification to Intersection Algorithm

The initial results from the contrast model reflected a lack of symmetry between document sets during intersection. This was corrected by implementing a function to assess intersection inspired by Tversky's equation presented in chapter two: "$f(A \cap B) = f(A \cap B) - \sum X_a X_b / N_x$, where the summation is over all X in $A \cap B$."

Tversky's algorithm was designed for the comparison of object's whose features are comparable at one time. The two objects in Figure 9.30 .are simple objects that each possess nine features. Of these, Ball has three color features and Flower has five color features. The intersection becomes the intersection of common features, for these objects their common feature is *color*.

Ball Features			Flower Features		
Color	(3)	3/9	Color	(5)	5/9
Bounce	(2)	2/9	Smell	(4)	4/9
Texture	(4)	4/9			

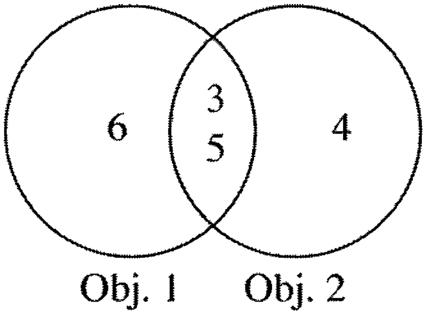


Figure 9.30 Intersection of Objects().

The equation in Figure 9.31 shows how Tversky's algorithm performs well at evaluating this intersection. But this algorithm needs a little tweak to work with multiple comparisons of multiple features.

$$\frac{\sum X_a X_b}{N_x} = \frac{\frac{1}{3} * \frac{5}{9}}{2} = \frac{4}{9}$$

Figure 9.31 Equation to Evaluate Intersection of Objects().

A slight variation on Tversky's algorithm has been created for the evaluation of document intersection. The equation is shown in Figure 9.32. X_a is a measurement of common edges, reflecting the proportion of path matches to path length totals in document *a* and X_b is the proportion of path matches to path length in document *b*. N_x is replaced with T_a T_b, the product of the total number of edges in document *a* and *b*.

$$\frac{\sum X_a X_b}{T_a T_b}$$

$$X = \frac{\#\ \textit{edge matches}}{\#\ \textit{edges per path}} \qquad T = \begin{matrix}\#\ \textit{total edges} \\ \textit{in a document}\end{matrix}$$

Figure 9.32 Equation to Evaluate Intersection of Document Paths.

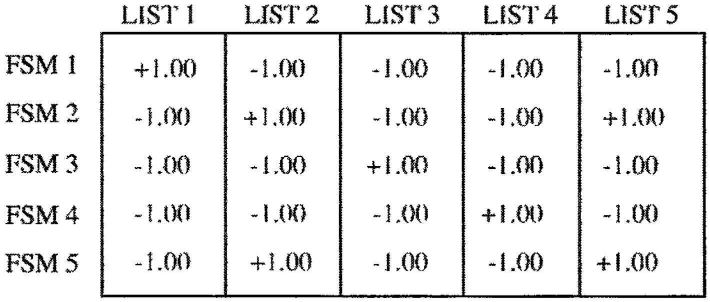
The pseudo-code for the added intersection function is shown in Figure 9.33. The values X_a and X_b are summed in *intersectDifference*(). The psudo-code for *intersectDifference*() is shown if Figure 8.35.

```
Pseudo-code for intersection()
{   for ( i =0; i < NUMDOCS; i++)
        for ( j =0; j< NUMDOCS; j++)
            intersection[i][j]=
            ( ( (sumSame/Edge [i][j] * sumSame/Edge [j][i] ) /
                ( edgeCount [i][j] * edgeCount [j][i] ) ) )
}
```

Figure 9.33 Pseudo-Code for Intersection().

9.7 Modifications to the Product Moment Calculation

The initial Product Moment algorithm began with the ratio of same edges to total edges for assessing matching deviation scores and the ratio of non-matching edges to total edges for non-matching deviation scores. This type of input results in an absolute negative correlation of path edges. Figure 9.34 demonstrates the scores produced by the initial product moment algorithm on the five documents presented in 9.4.



	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5
FSM 1	+1.00	-1.00	-1.00	-1.00	-1.00
FSM 2	-1.00	+1.00	-1.00	-1.00	+1.00
FSM 3	-1.00	-1.00	+1.00	-1.00	-1.00
FSM 4	-1.00	-1.00	-1.00	+1.00	-1.00
FSM 5	-1.00	+1.00	-1.00	-1.00	+1.00

Figure 9.34 Negative Correlation in Product Moment Calculation.

To resolve this issue, the counts of matching and non-matching edges are used as input. These counts reflect the distance of semantic overlap and separation. By removing the ratios, the calculations more closely represent semantic distance.

9.8 Future Evolution

The last state of an edge list becomes an accepting state in the finite state machine of a document. It reflects the synset where the search word is found. The next to the last state is the subsuming concept. Matches down to last node match specific word concepts. These are 'word' matches, not 'concept' matches To catch the entire synset that a state belongs to, the next to the last state could be the last counted edge.

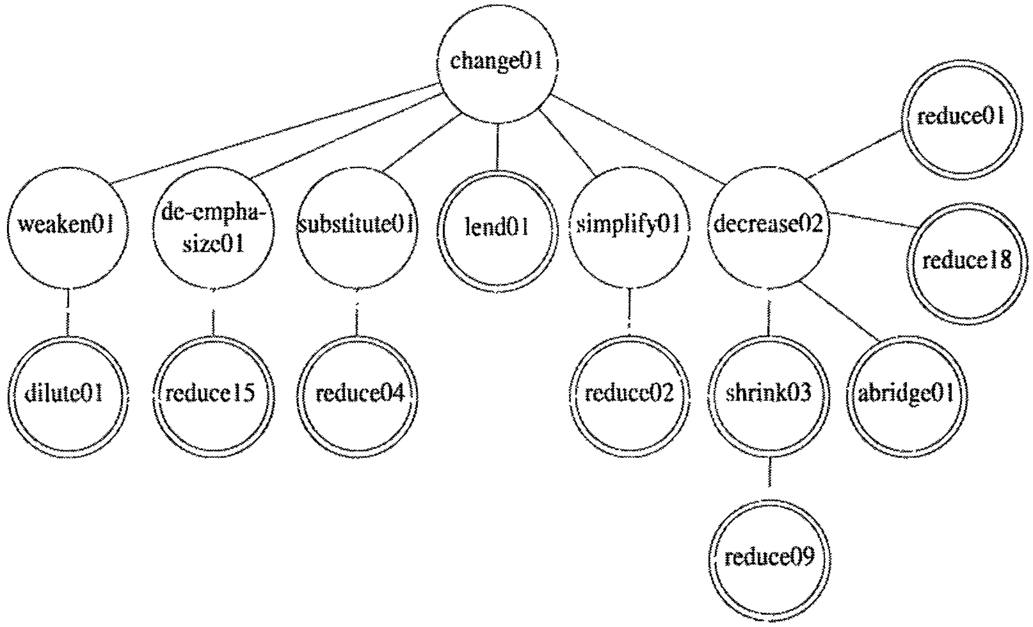


Figure 9.37 FSM Concept Subsumption.

Consider the synset example in Figure 9.37. There are four members of the synset decrease02. All of the next states for this state are accepting states. Three of the concepts have no children and could be universally referred to through the subsuming concept decrease02.

The effect of prunıng back accepting states in similarity detection is an area deemed beyond the scope of this project. It could provide an interestıng method of reducing a problem's complexity in future study.

CHAPTER 10

CONCLUSIONS

10.1 Schroedinger's Similarity

At any moment the relation of similarity exists in an infinite number of states. The properties of this relation are directly related to the properties searched for. That is, like the cat in the box, it is the act of observation which causes similarity to arise.

10.2 Similarity is in the Eye of the Beholder

The types of documents found to be similar for each data set given each algorithm were directly related to the types of similarities looked for. This seems to imply that similarity is not absolute, but relative to the observer.

In the results produced by the Contrast Model, searching for a literal semantic distance correlation produced shallow results of documents whose prefixes were similar in length. The Ratio Model looked for similarity based on shared proportions of edge lists. The resulting clusters of documents shared this trait. The Product Moment algorithm looked for correlation of path prefixes between documents. The results showed highly condensed clusters of documents containing the highest level of correlation.

These results seem consistent with how the human mind assesses similarity. Similarity is a relationship that is defined through comparisons of known objects possessing known features within a know framework. If there is no frame of reference, there can be no measurable similarity.

10.3 Forward Momentum

To closer align the results of similarity assessment with the expectations of a human observer, an exact definition of what similarity *is,* is required. This may be a vain pursuit as few human observers are likely to define similarity in exactly the same way.

The definition of similarity drives the results. There is an adage, 'You can please some of the people all of the time and all of the people some of the time, but you can't please all of the people all of the time.' This applies to the definition of similarity. Even though it may not be possible to observe detect all types of similarity at the same moment, in the same observation. It is certainly possible to detect similarity.

The question of whether similarity can be inferred from a set of documents has transitioned to can similarity more closely reflect the goals of the observer. The answer is yes. The impetus lies in identifying precisely what type of similarity is desired.

Once you know what you're looking for, you can find it.

APPENDIX A

DATA SET 1 IDENTIFICATION NUMBERS

NEWID refers to Reuters-21578 Document Identification Number.
DS1ID refers to Data Set 1 Document Identification Number.

Data Set 1:
NEWID = 1, DS1ID = 1
NEWID = 2, DS1ID = 2
NEWID = 136, DS1ID = 3
NEWID = 7, DS1ID = 4
NEWID = 8, DS1ID = 5
NEWID = 9, DS1ID = 6
NEWID = 10, DS1ID = 7
NEWID = 12, DS1ID = 8
NEWID = 15, DS1ID = 9
NEWID = 16, DS1ID = 10
NEWID = 17, DS1ID = 11
NEWID = 18, DS1ID = 12
NEWID = 19, DS1ID = 13
NEWID = 20, DS1ID = 14
NEWID = 21, DS1ID = 15
NEWID = 23, DS1ID = 16
NEWID = 26, DS1ID = 17
NEWID = 29, DS1ID = 18
NEWID = 32, DS1ID = 19
NEWID = 34, DS1ID = 20
NEWID = 39, DS1ID = 21
NEWID = 40, DS1ID = 22
NEWID = 42, DS1ID = 23
NEWID = 43, DS1ID = 24
NEWID = 44, DS1ID = 25
NEWID = 45, DS1ID = 26
NEWID = 46, DS1ID = 27
NEWID = 47, DS1ID = 28
NEWID = 51, DS1ID = 29
NEWID = 54, DS1ID = 30
NEWID = 55, DS1ID = 31
NEWID = 56, DS1ID = 32
NEWID = 57, DS1ID = 33
NEWID = 60, DS1ID = 34
NEWID = 61, DS1ID = 35
NEWID = 62, DS1ID = 36
NEWID = 63, DS1ID = 37
NEWID = 67, DS1ID = 38
NEWID = 70, DS1ID = 39
NEWID = 75, DS1ID = 40
NEWID = 79, DS1ID = 41
NEWID = 80, DS1ID = 42
NEWID = 81, DS1ID = 43
NEWID = 84, DS1ID = 44
NEWID = 88, DS1ID = 45
NEWID = 90, DS1ID = 46
NEWID = 91, DS1ID = 47
NEWID = 96, DS1ID = 48
NEWID = 100, DS1ID = 49
NEWID = 107, DS1ID = 50

APPENDIX B

DATA SET 2 IDENTIFICATION NUMBERS

NEWID refers to Reuters-21578 Document Identification Number.
DS2ID refers to Data Set 2 Document Identification Number.

Data Set 2:
NEWID = 110, DS2ID = 1
NEWID = 111, DS2ID = 2
NEWID = 112, DS2ID = 3
NEWID = 113, DS2ID = 4
NEWID = 114, DS2ID = 5
NEWID = 115, DS2ID = 6
NEWID = 116, DS2ID = 7
NEWID = 117, DS2ID = 8
NEWID = 118, DS2ID = 9
NEWID = 121, DS2ID = 10
NEWID = 123, DS2ID = 11
NEWID = 125, DS2ID = 12
NEWID = 127, DS2ID = 13
NEWID = 130, DS2ID = 14
NEWID = 131, DS2ID = 15
NEWID = 135, DS2ID = 16
NEWID = 136, DS2ID = 17
NEWID = 137, DS2ID = 18
NEWID = 141, DS2ID = 19
NEWID = 144, DS2ID = 20
NEWID = 145, DS2ID = 21
NEWID = 147, DS2ID = 22
NEWID = 149, DS2ID = 23
NEWID = 153, DS2ID = 24
NEWID = 157, DS2ID = 25
NEWID = 172, DS2ID = 26
NEWID = 175, DS2ID = 27
NEWID = 176, DS2ID = 28
NEWID = 177, DS2ID = 29
NEWID = 178, DS2ID = 30
NEWID = 179, DS2ID = 31
NEWID = 180, DS2ID = 32
NEWID = 181, DS2ID = 33
NEWID = 190, DS2ID = 34
NEWID = 192, DS2ID = 35
NEWID = 193, DS2ID = 36
NEWID = 195, DS2ID = 37
NEWID = 197, DS2ID = 38
NEWID = 200, DS2ID = 39
NEWID = 203, DS2ID = 40
NEWID = 205, DS2ID = 41
NEWID = 208, DS2ID = 42
NEWID = 209, DS2ID = 43
NEWID = 218, DS2ID = 44
NEWID = 219, DS2ID = 45
NEWID = 223, DS2ID = 46
NEWID = 225, DS2ID = 47
NEWID = 227, DS2ID = 48
NEWID = 230, DS2ID = 49
NEWID = 232, DS2ID = 50

APPENDIX C

DATA SET 3 IDENTIFICATION NUMBERS

NEWID refers to Reuters-21578 Document Identification Number.
DS3ID refers to Data Set 3 Document Identification Number.

Data Set 3:

NEWID = 234, DS3ID = 1
NEWID = 235, DS3ID = 2
NEWID = 236, DS3ID = 3
NEWID = 237, DS3ID = 4
NEWID = 238, DS3ID = 5
NEWID = 240, DS3ID = 6
NEWID = 241, DS3ID = 7
NEWID = 243, DS3ID = 8
NEWID = 246, DS3ID = 9
NEWID = 247, DS3ID = 10
NEWID = 248, DS3ID = 11
NEWID = 249, DS3ID = 12
NEWID = 252, DS3ID = 13
NEWID = 256, DS3ID = 14
NEWID = 257, DS3ID = 15
NEWID = 259, DS3ID = 16
NEWID = 263, DS3ID = 17
NEWID = 264, DS3ID = 18
NEWID = 265, DS3ID = 19
NEWID = 266, DS3ID = 20
NEWID = 269, DS3ID = 21
NEWID = 270, DS3ID = 22
NEWID = 273, DS3ID = 23
NEWID = 274, DS3ID = 24
NEWID = 275, DS3ID = 25
NEWID = 276, DS3ID = 26
NEWID = 278, DS3ID = 27
NEWID = 279, DS3ID = 28
NEWID = 288, DS3ID = 29
NEWID = 292, DS3ID = 30
NEWID = 295, DS3ID = 31
NEWID = 297, DS3ID = 32
NEWID = 300, DS3ID = 33
NEWID = 302, DS3ID = 34
NEWID = 303, DS3ID = 35
NEWID = 304, DS3ID = 36
NEWID = 305, DS3ID = 37
NEWID = 306, DS3ID = 38
NEWID = 308, DS3ID = 39
NEWID = 309, DS3ID = 40
NEWID = 311, DS3ID = 41
NEWID = 316, DS3ID = 42
NEWID = 317, DS3ID = 43
NEWID = 319, DS3ID = 44
NEWID = 320, DS3ID = 45
NEWID = 331, DS3ID = 46
NEWID = 335, DS3ID = 47
NEWID = 337, DS3ID = 48
NEWID = 339, DS3ID = 49
NEWID = 340, DS3ID = 50

APPENDIX D

DATA SET 4 IDENTIFICATION NUMBERS

NEWID refers to Reuters-21578 Document Identification Number.
DS4ID refers to Data Set 4 Document Identification Number.

Data Set 4:

NEWID = 342, DS4ID = 1
NEWID = 346, DS4ID = 2
NEWID = 347, DS4ID = 3
NEWID = 348, DS4ID = 4
NEWID = 350, DS4ID = 5
NEWID = 354, DS4ID = 6
NEWID = 355, DS4ID = 7
NEWID = 356, DS4ID = 8
NEWID = 357, DS4ID = 9
NEWID = 358, DS4ID = 10
NEWID = 359, DS4ID = 11
NEWID = 362, DS4ID = 12
NEWID = 367, DS4ID = 13
NEWID = 370, DS4ID = 14
NEWID = 372, DS4ID = 15
NEWID = 375, DS4ID = 16
NEWID = 382, DS4ID = 17
NEWID = 386, DS4ID = 18
NEWID = 390, DS4ID = 19
NEWID = 393, DS4ID = 20
NEWID = 401, DS4ID = 21
NEWID = 419, DS4ID = 22
NEWID = 432, DS4ID = 23
NEWID = 434, DS4ID = 24
NEWID = 435, DS4ID = 25
NEWID = 448, DS4ID = 26
NEWID = 454, DS4ID = 27
NEWID = 458, DS4ID = 28
NEWID = 473, DS4ID = 29
NEWID = 479, DS4ID = 30
NEWID = 481, DS4ID = 31
NEWID = 482, DS4ID = 32
NEWID = 489, DS4ID = 33
NEWID = 491, DS4ID = 34
NEWID = 495, DS4ID = 35
NEWID = 496, DS4ID = 36
NEWID = 500, DS4ID = 37
NEWID = 501, DS4ID = 38
NEWID = 502, DS4ID = 39
NEWID = 503, DS4ID = 40
NEWID = 504, DS4ID = 41
NEWID = 505, DS4ID = 42
NEWID = 507, DS4ID = 43
NEWID = 515, DS4ID = 44
NEWID = 516, DS4ID = 45
NEWID = 518, DS4ID = 46
NEWID = 523, DS4ID = 47
NEWID = 524, DS4ID = 48
NEWID = 525, DS4ID = 49
NEWID = 540, DS4ID = 50

APPENDIX E

DATA SET 5 IDENTIFICATION NUMBERS

NEWID refers to Reuters-21578 Document Identification Number.
DS5ID refers to Data Set 5 Document Identification Number.

Data Set 5:

NEWID = 544, DS5ID = 1
NEWID = 545, DS5ID = 2
NEWID = 549, DS5ID = 3
NEWID = 550, DS5ID = 4
NEWID = 551, DS5ID = 5
NEWID = 561, DS5ID = 6
NEWID = 562, DS5ID = 7
NEWID = 563, DS5ID = 8
NEWID = 564, DS5ID = 9
NEWID = 565, DS5ID = 10
NEWID = 566, DS5ID = 11
NEWID = 567, DS5ID = 12
NEWID = 568, DS5ID = 13
NEWID = 570, DS5ID = 14
NEWID = 582, DS5ID = 15
NEWID = 583, DS5ID = 16
NEWID = 597, DS5ID = 17
NEWID = 627, DS5ID = 18
NEWID = 629, DS5ID = 19
NEWID = 634, DS5ID = 20
NEWID = 635, DS5ID = 21
NEWID = 637, DS5ID = 22
NEWID = 654, DS5ID = 23
NEWID = 659, DS5ID = 24
NEWID = 660, DS5ID = 25
NEWID = 668, DS5ID = 26
NEWID = 669, DS5ID = 27
NEWID = 671, DS5ID = 28
NEWID = 672, DS5ID = 29
NEWID = 674, DS5ID = 30
NEWID = 685, DS5ID = 31
NEWID = 696, DS5ID = 32
NEWID = 704, DS5ID = 33
NEWID = 714, DS5ID = 34
NEWID = 677, DS5ID = 35
NEWID = 718, DS5ID = 36
NEWID = 721, DS5ID = 37
NEWID = 730, DS5ID = 38
NEWID = 732, DS5ID = 39
NEWID = 740, DS5ID = 40
NEWID = 742, DS5ID = 41
NEWID = 743, DS5ID = 42
NEWID = 748, DS5ID = 43
NEWID = 749, DS5ID = 44
NEWID = 759, DS5ID = 45
NEWID = 768, DS5ID = 46
NEWID = 769, DS5ID = 47
NEWID = 770, DS5ID = 48
NEWID = 798, DS5ID = 49
NEWID = 799, DS5ID = 50

APPENDIX F

CONTRAST MODEL RESULTS

DOCUMENT SETS 2-5

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.16	-0.20	-0.08	-0.23	-0.01	0.14	-0.18	-0.22	-0.12
FSM 2	-0.33	1.00	-0.56	-0.34	-0.51	-0.35	-0.36	-0.47	-0.47	-0.33
FSM 3	-0.28	-0.46	1.00	-0.40	-0.32	-0.40	-0.33	-0.48	-0.41	-0.37
FSM 4	-0.32	-0.42	-0.56	1.00	-0.49	-0.23	-0.24	-0.36	-0.43	-0.39
FSM 5	-0.37	-0.52	-0.46	-0.43	1.00	-0.39	-0.43	-0.47	-0.47	-0.42
FSM 6	-0.33	-0.46	-0.59	-0.26	-0.49	1.00	-0.33	-0.45	-0.49	-0.34
FSM 7	-0.20	-0.41	-0.52	-0.24	-0.47	-0.29	1.00	-0.16	-0.25	-0.37
FSM 8	-0.39	-0.53	-0.61	-0.32	-0.52	-0.35	-0.12	1.00	-0.48	-0.50
FSM 9	-0.32	-0.45	-0.50	-0.31	-0.43	-0.38	-0.13	-0.42	1.00	-0.38
FSM 10	-0.21	-0.30	-0.44	-0.28	-0.39	-0.17	-0.26	-0.42	-0.39	1.00

Figure F.1 Data Set 2 – Contrast Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1							■			
2	■									
3	■									
4						■				
5	■									
6				■						
7								■		
8							■			
9							■			
10						■				

Figure F.2 Data Set 2 – Contrast Model – Similarity Grid.

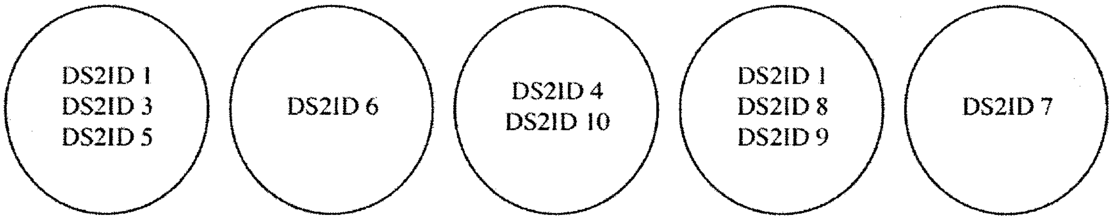


Figure F.3 Data Set 2 – Contrast Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.35	-0.29	-0.23	-0.24	-0.26	-0.41	-0.35	-0.20	-0.23
FSM 2	-0.22	1.00	-0.11	0.07	-0.09	-0.15	-0.05	-0.31	0.01	-0.04
FSM 3	-0.17	-0.11	1.00	-0.05	-0.03	-0.09	-0.23	-0.24	0.05	0.04
FSM 4	-0.05	0.13	-0.01	1.00	0.00	-0.03	-0.19	-0.18	0.07	0.06
FSM 5	-0.09	-0.05	0.02	0.01	1.00	-0.09	-0.20	-0.21	0.07	0.07
FSM 6	-0.07	-0.08	-0.04	-0.02	-0.07	1.00	-0.20	-0.20	-0.02	0.08
FSM 7	-0.33	-0.19	-0.30	-0.31	-0.31	-0.34	1.00	-0.39	-0.25	-0.19
FSM 8	-0.34	-0.37	-0.35	-0.31	-0.33	-0.39	-0.45	1.00	-0.33	-0.29
FSM 9	-0.11	-0.03	0.02	-0.02	-0.01	-0.10	-0.22	-0.29	1.00	-0.04
FSM 10	-0.07	0.03	0.12	0.09	0.10	0.10	0.03	-0.13	0.05	1.00

Figure F.4 Data Set 3 – Contrast Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1									■	
2				■						
3									■	
4		■								
5										■
6										■
7		■								
8										■
9			■							
10			■							

Figure F.5 Data Set 3 – Contrast Model – Similarity Grid.

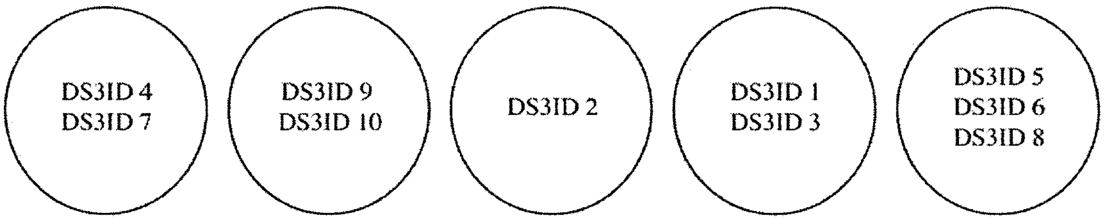


Figure F.6 Data Set 3 – Contrast Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.03	0.05	0.07	0.14	0.00	0.15	0.11	0.08	0.01
FSM 2	-0.07	1.00	-0.21	-0.11	-0.18	-0.06	-0.19	-0.20	-0.21	-0.08
FSM 3	0.00	-0.17	1.00	-0.04	0.12	-0.11	0.25	0.17	-0.09	-0.08
FSM 4	-0.08	-0.13	-0.11	1.00	-0.04	-0.16	-0.19	-0.12	-0.28	0.01
FSM 5	0.11	-0.13	0.16	0.09	1.00	-0.03	0.09	0.09	0.01	-0.02
FSM 6	-0.08	-0.07	-0.16	-0.15	-0.11	1.00	-0.15	-0.09	-0.21	-0.11
FSM 7	-0.03	-0.24	0.07	-0.23	-0.06	-0.20	1.00	0.18	-0.21	-0.10
FSM 8	0.03	-0.19	0.13	-0.08	0.02	-0.07	0.32	1.00	-0.08	0.04
FSM 9	-0.08	-0.26	-0.18	-0.28	-0.11	-0.22	-0.19	-0.15	1.00	-0.13
FSM 10	-0.16	-0.18	-0.21	-0.07	-0.17	-0.20	-0.15	-0.08	-0.20	1.00

Figure F.7 Data Set 4 – Contrast Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1							■			
2						■				
3							■			
4										■
5			■							
6		■								
7								■		
8							■			
9	■									
10				■						

Figure F.8 Data Set 4 – Contrast Model – Similarity Grid.

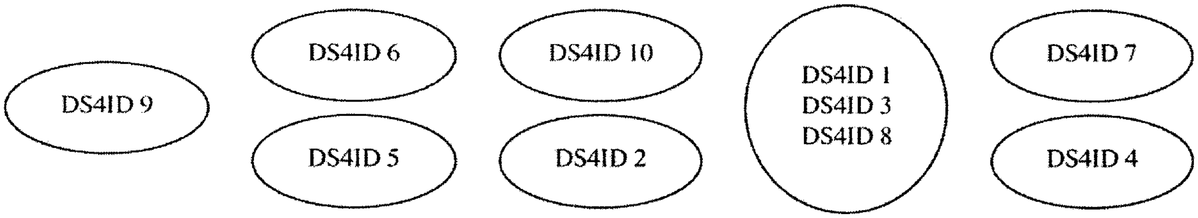


Figure F.9 Data Set 4 – Contrast Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.31	-0.31	-0.32	-0.29	-0.36	-0.36	-0.28	-0.32	-0.28
FSM 2	-0.26	1.00	-0.25	-0.26	-0.20	-0.22	-0.19	-0.07	-0.23	-0.02
FSM 3	-0.25	-0.21	1.00	-0.27	-0.01	-0.29	-0.24	-0.13	-0.26	-0.22
FSM 4	-0.38	-0.29	-0.35	1.00	-0.10	-0.28	-0.41	-0.21	-0.30	-0.25
FSM 5	-0.22	-0.17	-0.04	0.05	1.00	-0.30	-0.22	-0.13	-0.19	-0.21
FSM 6	-0.46	-0.36	-0.46	-0.39	-0.49	1.00	-0.45	-0.34	-0.38	-0.37
FSM 7	-0.37	-0.21	-0.29	-0.33	-0.27	-0.32	1.00	-0.21	-0.34	-0.27
FSM 8	-0.30	-0.12	-0.21	-0.17	-0.22	-0.24	-0.22	1.00	-0.24	-0.15
FSM 9	-0.38	-0.27	-0.34	-0.28	-0.29	-0.31	-0.38	-0.25	1.00	-0.25
FSM 10	-0.33	-0.15	-0.31	-0.25	-0.30	-0.28	-0.30	-0.20	-0.26	1.00

Figure F.10 Data Set 5 – Contrast Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1								■		
2										■
3					■					
4					■					
5				■						
6								■		
7								■		
8		■								
9										■
10		■								

Figure F.11 Data Set 5 – Contrast Model – Similarity Grid.

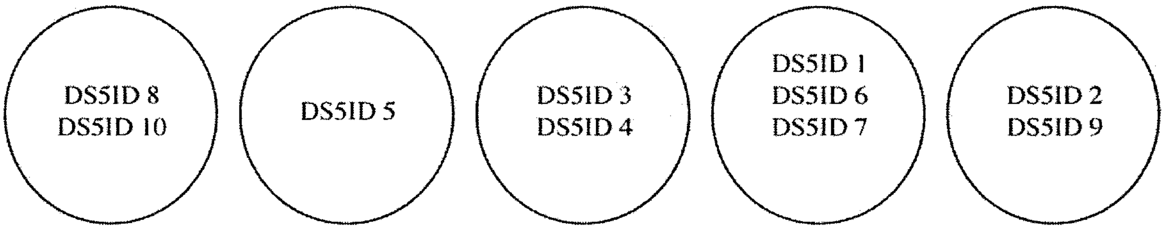


Figure F.12 Data Set 5 – Contrast Model – Clusters of Similar Documents.

APPENDIX G

RATIO MODEL RESULTS

DOCUMENT SETS 2-5

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.11	0.13	0.10	0.08	0.11	0.16	0.07	0.11	0.23
FSM 2	0.40	1.00	0.22	0.27	0.20	0.27	0.30	0.20	0.25	0.37
FSM 3	0.43	0.21	1.00	0.21	0.24	0.22	0.27	0.16	0.24	0.34
FSM 4	0.40	0.28	0.23	1.00	0.23	0.35	0.37	0.27	0.28	0.35
FSM 5	0.39	0.23	0.28	0.25	1.00	0.28	0.28	0.21	0.26	0.34
FSM 6	0.38	0.25	0.21	0.32	0.23	1.00	0.32	0.22	0.24	0.37
FSM 7	0.44	0.25	0.23	0.30	0.21	0.29	1.00	0.31	0.33	0.34
FSM 8	0.37	0.24	0.21	0.32	0.23	0.32	0.45	1.00	0.27	0.30
FSM 9	0.41	0.24	0.24	0.27	0.23	0.25	0.39	0.20	1.00	0.35
FSM 10	0.44	0.20	0.19	0.18	0.15	0.22	0.23	0.12	0.19	1.00

Figure G.1 Data Set 2 – Ratio Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1										■
2	■									
3	■									
4	■									
5	■									
6	■									
7	■									
8							■			
9	■									
10	■									

Figure G.2 Data Set 2 – Ratio Model – Similarity Grid.

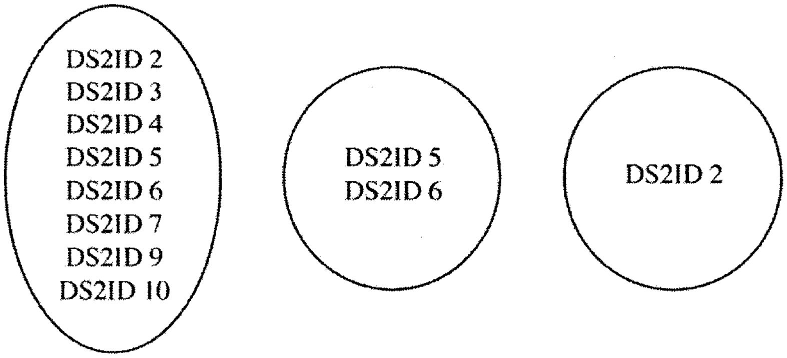


Figure G.3 Data Set 2 – Ratio Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.35	0.37	0.42	0.41	0.39	0.22	0.24	0.41	0.42
FSM 2	0.25	1.00	0.39	0.51	0.44	0.40	0.23	0.17	0.42	0.48
FSM 3	0.27	0.40	1.00	0.46	0.47	0.43	0.19	0.18	0.45	0.52
FSM 4	0.26	0.46	0.40	1.00	0.45	0.42	0.16	0.16	0.40	0.50
FSM 5	0.24	0.38	0.41	0.45	1.00	0.39	0.15	0.15	0.40	0.50
FSM 6	0.27	0.39	0.40	0.45	0.43	1.00	0.17	0.17	0.38	0.52
FSM 7	0.35	0.45	0.40	0.41	0.40	0.38	1.00	0.28	0.42	0.45
FSM 8	0.34	0.36	0.37	0.40	0.40	0.35	0.24	1.00	0.38	0.42
FSM 9	0.32	0.46	0.48	0.49	0.49	0.44	0.21	0.19	1.00	0.49
FSM 10	0.22	0.38	0.42	0.46	0.46	0.45	0.17	0.15	0.36	1.00

Figure G.4 Data Set 3 – Ratio Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1										■
2				■						
3										■
4										■
5										■
6										■
7										■
8										■
9										■
10					■					

Figure G.5 Data Set 3 – Ratio Model – Similarity Grid.

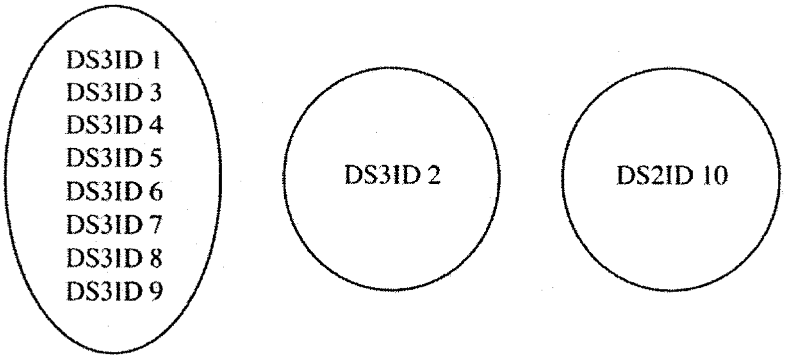


Figure G.6 Data Set 3 – Ratio Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.32	0.42	0.32	0.49	0.33	0.32	0.42	0.29	0.28
FSM 2	0.49	1.00	0.41	0.39	0.43	0.44	0.34	0.40	0.32	0.39
FSM 3	0.49	0.29	1.00	0.34	0.53	0.34	0.42	0.51	0.28	0.31
FSM 4	0.49	0.40	0.46	1.00	0.50	0.40	0.34	0.44	0.30	0.44
FSM 5	0.52	0.29	0.50	0.36	1.00	0.35	0.33	0.44	0.29	0.30
FSM 6	0.48	0.40	0.42	0.35	0.45	1.00	0.33	0.44	0.30	0.36
FSM 7	0.52	0.36	0.55	0.36	0.49	0.39	1.00	0.59	0.35	0.41
FSM 8	0.51	0.31	0.53	0.34	0.49	0.38	0.49	1.00	0.31	0.38
FSM 9	0.50	0.35	0.44	0.34	0.47	0.39	0.36	0.44	1.00	0.41
FSM 10	0.45	0.38	0.41	0.42	0.44	0.38	0.37	0.46	0.34	1.00

Figure G.7 Data Set 4 – Ratio Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1					■					
2	■									
3					■					
4					■					
5	■									
6	■									
7								■		
8			■							
9	■									
10								■		

Figure G.8 Data Set 4 – Ratio Model – Similarity Grid.

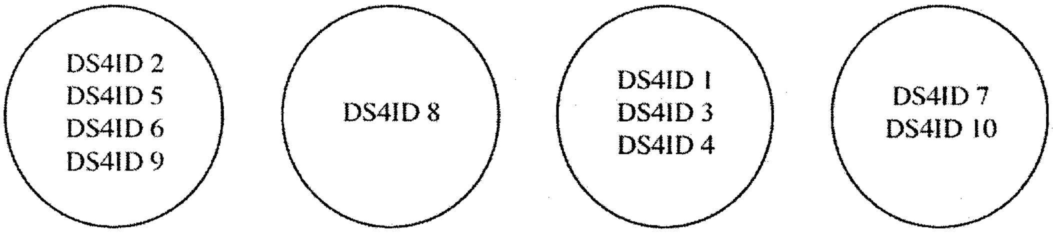


Figure G.9 Data Set 4 – Ratio Model – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.38	0.39	0.34	0.39	0.25	0.34	0.41	0.36	0.33
FSM 2	0.21	1.00	0.34	0.26	0.34	0.18	0.31	0.42	0.29	0.30
FSM 3	0.20	0.34	1.00	0.24	0.42	0.15	0.28	0.38	0.27	0.22
FSM 4	0.24	0.37	0.34	1.00	0.45	0.24	0.28	0.41	0.33	0.30
FSM 5	0.24	0.38	0.45	0.39	1.00	0.17	0.31	0.40	0.32	0.25
FSM 6	0.26	0.37	0.32	0.32	0.29	1.00	0.30	0.38	0.34	0.31
FSM 7	0.22	0.39	0.36	0.27	0.35	0.20	1.00	0.40	0.29	0.27
FSM 8	0.20	0.39	0.36	0.28	0.34	0.17	0.30	1.00	0.28	0.26
FSM 9	0.22	0.36	0.34	0.30	0.35	0.21	0.28	0.38	1.00	0.28
FSM 10	0.29	0.45	0.38	0.36	0.37	0.27	0.36	0.43	0.37	1.00

Figure G.10 Data Set 5 – Ratio Model – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1								■		
2								■		
3					■					
4					■					
5			■							
6								■		
7								■		
8		■								
9								■		
10		■								

Figure G.11 Data Set 5 – Ratio Model – Similarity Grid.

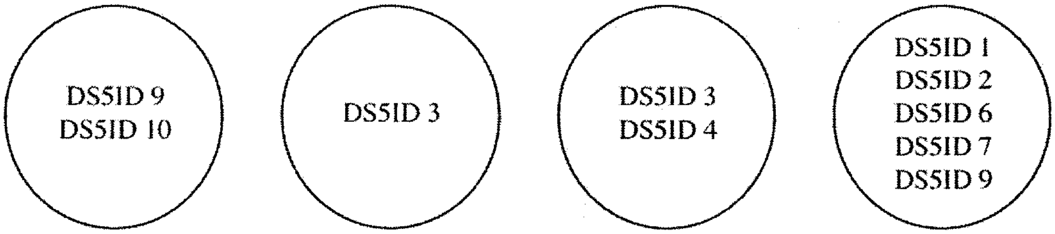


Figure G.12 Data Set 5 – Ratio Model – Clusters of Similar Documents.

APPENDIX H

PRODUCT MOMENT RESULTS

DOCUMENT SETS 2-5

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.31	-0.15	-0.29	-0.00	-0.34	-0.37	0.00	-0.34	0.05
FSM 2	0.06	1.00	0.15	0.16	0.29	-0.07	0.15	0.23	-0.03	0.05
FSM 3	0.23	0.26	1.00	0.20	0.25	0.20	0.44	0.31	-0.07	0.25
FSM 4	0.03	-0.09	0.10	1.00	0.21	-0.09	-0.16	-0.20	0.01	0.15
FSM 5	0.15	0.05	0.12	0.15	1.00	0.01	0.13	-0.00	-0.17	0.19
FSM 6	-0.05	-0.32	0.20	-0.04	0.11	1.00	0.06	0.00	-0.08	0.04
FSM 7	-0.12	-0.15	0.13	-0.25	0.22	0.12	1.00	-0.49	-0.06	0.08
FSM 8	0.11	-0.14	0.19	-0.10	0.06	0.09	-0.30	1.00	-0.12	0.12
FSM 9	0.10	-0.03	0.05	0.05	-0.01	-0.03	0.15	0.04	1.00	0.16
FSM 10	0.06	-0.33	0.07	0.12	0.08	-0.23	0.02	0.01	-0.12	1.00

Figure H.1 Data Set 2 – Product Moment – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1										■
2					■					
3							■			
4					■					
5										■
6			■							
7					■					
8			■							
9										■
10				■						

Figure H.2 Data Set 2 – Product Moment – Similarity Grid.

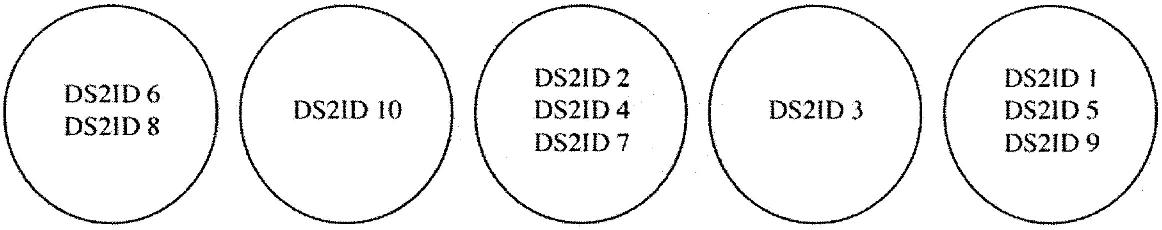


Figure H.3 Data Set 2 – Product Moment – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.12	0.03	0.02	-0.04	-0.11	-0.07	0.16	-0.01	0.05
FSM 2	-0.02	1.00	-0.08	-0.20	-0.16	-0.09	-0.34	0.20	-0.23	-0.19
FSM 3	0.07	0.04	1.00	-0.14	-0.09	-0.06	0.06	0.19	-0.23	-0.08
FSM 4	-0.00	-0.07	-0.19	1.00	-0.26	-0.14	-0.11	0.21	-0.18	-0.22
FSM 5	-0.15	0.00	-0.13	-0.23	1.00	-0.11	0.09	-0.04	-0.15	-0.19
FSM 6	-0.07	0.06	-0.04	-0.07	-0.06	1.00	0.13	0.15	-0.08	-0.00
FSM 7	0.13	0.02	0.12	0.09	0.18	0.20	1.00	0.19	0.08	0.14
FSM 8	0.07	0.30	0.11	0.14	0.05	0.09	-0.08	1.00	0.01	0.12
FSM 9	0.05	0.01	-0.14	-0.11	-0.12	-0.08	0.01	0.11	1.00	-0.11
FSM 10	0.03	-0.13	-0.14	-0.30	-0.29	-0.08	-0.06	0.16	-0.26	1.00

Figure H.4 Data Set 3 – Product Moment – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1								■		
2								■		
3								■		
4								■		
5							■			
6								■		
7						■				
8		■								
9								■		
10								■		

Figure H.5 Data Set 3 – Product Moment – Similarity Grid.

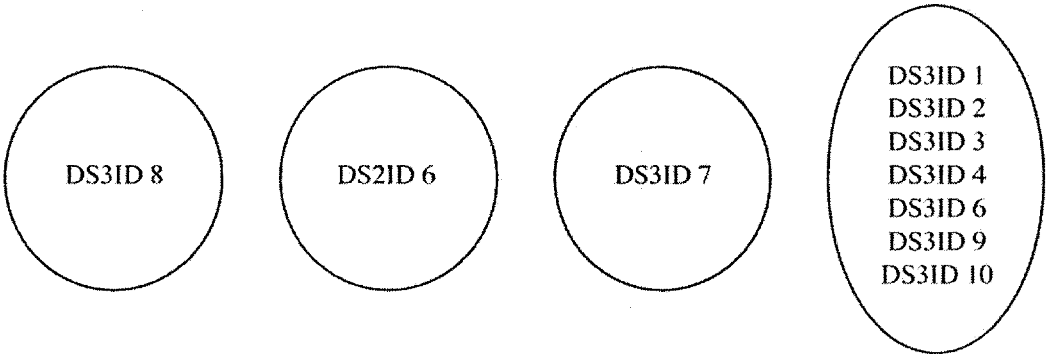


Figure H.6 Data Set 3 – Product Moment – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	-0.22	-0.12	-0.01	-0.12	-0.04	-0.11	-0.15	-0.09	-0.03
FSM 2	-0.01	1.00	0.05	0.12	0.09	0.10	-0.04	-0.03	0.03	-0.06
FSM 3	-0.05	-0.18	1.00	-0.21	-0.26	-0.02	0.06	-0.01	-0.08	-0.03
FSM 4	-0.03	-0.07	-0.26	1.00	-0.19	-0.12	-0.00	-0.08	-0.07	0.02
FSM 5	-0.05	-0.10	-0.26	-0.11	1.00	-0.05	-0.07	-0.12	-0.14	-0.03
FSM 6	-0.04	-0.13	-0.07	-0.11	-0.12	1.00	-0.12	-0.17	-0.08	0.01
FSM 7	-0.02	-0.11	0.04	0.15	-0.05	0.05	1.00	-0.16	0.14	-0.05
FSM 8	-0.09	-0.25	0.00	0.03	-0.13	-0.11	-0.25	1.00	0.00	-0.01
FSM 9	0.04	-0.03	0.01	0.07	-0.09	0.04	0.07	0.04	1.00	0.02
FSM 10	0.01	-0.23	0.04	0.08	-0.04	0.08	-0.13	-0.03	0.04	1.00

Figure H.7 Data Set 4– Product Moment – Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1				■						
2				■						
3							■			
4										■
5										■
6										■
7				■						
8				■						
9							■			
10						■				

Figure H.8 Data Set 4 – Product Moment – Similarity Grid.

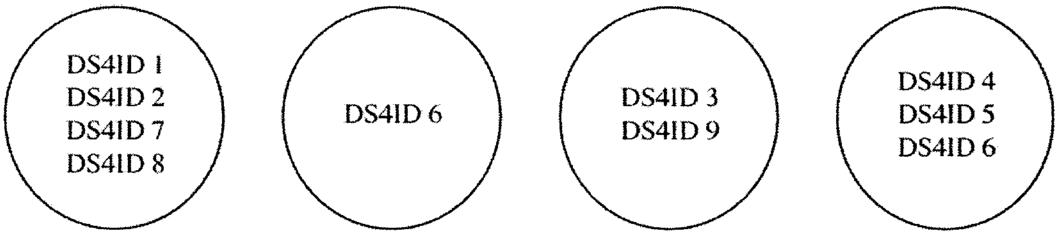


Figure H.9 Data Set 4 – Product Moment – Clusters of Similar Documents.

	LIST 1	LIST 2	LIST 3	LIST 4	LIST 5	LIST 6	LIST 7	LIST 8	LIST 9	LIST 10
FSM 1	1.00	0.10	0.17	0.07	0.06	0.31	0.13	0.02	0.06	-0.02
FSM 2	-0.42	1.00	0.08	-0.25	-0.37	-0.27	-0.23	-0.12	-0.16	-0.49
FSM 3	-0.26	0.11	1.00	-0.05	0.14	0.27	-0.11	-0.09	0.03	-0.18
FSM 4	-0.15	0.09	0.10	1.00	-0.05	0.07	0.06	-0.23	-0.11	-0.15
FSM 5	-0.30	-0.13	0.16	-0.01	1.00	0.11	-0.04	-0.08	-0.03	-0.19
FSM 6	-0.05	0.01	0.13	-0.07	-0.01	1.00	0.13	-0.08	-0.02	-0.16
FSM 7	-0.06	-0.05	-0.05	0.00	-0.03	0.23	1.00	-0.17	0.02	-0.15
FSM 8	-0.10	0.04	0.08	-0.26	-0.07	0.29	-0.01	1.00	0.07	-0.23
FSM 9	-0.34	-0.02	0.04	-0.16	-0.03	-0.09	-0.01	-0.12	1.00	-0.23
FSM 10	-0.08	-0.09	0.16	-0.04	0.08	0.14	0.14	-0.09	0.02	1.00

Figure H.10 Data Set 5 – Product Moment– Result Matrix.

	1	2	3	4	5	6	7	8	9	10
1						■				
2			■							
3						■				
4			■							
5			■							
6			■							
7						■				
8						■				
9			■							
10			■							

Figure H.11 Data Set 5 – Product Moment – Similarity Grid.

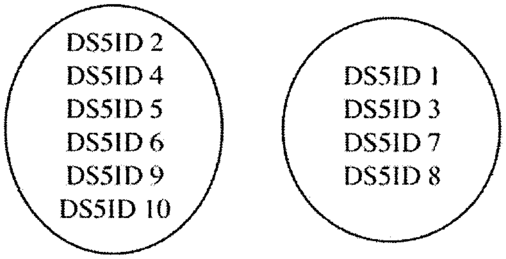


Figure H.12 Data Set 5 – Product Moment – Clusters of Similar Documents.

APPENDIX I

CONTRAST MODEL EXTENDED RESULTS

DATA SETS 1-5

	List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
FSM 1	1.00	-0.22	-0.15	-0.14	-0.14	-0.14	0.18	-0.05	-0.05	-0.10
FSM 2	-0.38	1.00	-0.46	-0.28	-0.29	-0.27	-0.27	-0.27	-0.15	-0.45
FSM 3	-0.21	-0.37	1.00	-0.27	-0.20	-0.21	-0.13	-0.14	-0.17	-0.23
FSM 4	-0.27	-0.25	-0.33	1.00	-0.16	-0.23	-0.20	-0.13	-0.14	-0.30
FSM 5	-0.30	-0.33	-0.32	-0.22	1.00	0.03	-0.07	0.08	-0.22	-0.31
FSM 6	-0.38	-0.40	-0.43	-0.36	-0.08	1.00	-0.21	0.10	-0.33	-0.43
FSM 7	0.07	-0.21	-0.17	-0.14	0.07	0.05	1.00	0.22	0.06	-0.11
FSM 8	-0.27	-0.35	-0.31	-0.22	0.03	0.20	0.01	1.00	-0.24	-0.31
FSM 9	-0.19	-0.04	-0.24	-0.10	-0.13	-0.16	-0.01	-0.13	1.00	-0.28
FSM 10	-0.02	-0.25	-0.07	-0.03	-0.01	-0.05	0.08	0.05	-0.08	1.00
FSM 11	-0.31	-0.06	-0.32	-0.27	-0.22	-0.21	-0.20	-0.11	-0.12	-0.33
FSM 12	-0.16	-0.24	-0.23	-0.13	-0.13	-0.09	-0.08	0.07	-0.18	-0.32
FSM 13	-0.32	-0.48	-0.35	0.02	-0.28	-0.35	-0.18	-0.20	-0.29	-0.39
FSM 14	-0.25	-0.33	-0.27	-0.22	-0.20	-0.21	-0.16	-0.05	-0.22	-0.32
FSM 15	-0.37	-0.35	-0.39	-0.36	-0.18	-0.19	-0.26	-0.20	-0.35	-0.50
FSM 16	-0.12	-0.24	-0.19	-0.13	0.02	0.39	0.04	0.15	-0.10	-0.18
FSM 17	0.04	-0.14	-0.11	-0.12	-0.09	-0.18	0.14	-0.04	-0.07	-0.13
FSM 18	-0.29	-0.45	-0.29	-0.32	-0.27	-0.28	-0.20	-0.18	-0.25	-0.34
FSM 19	-0.16	-0.19	-0.23	-0.15	-0.16	-0.17	-0.05	-0.12	-0.14	-0.24
FSM 20	-0.41	-0.33	-0.40	-0.35	-0.09	0.01	-0.18	-0.02	-0.11	-0.33
FSM 21	-0.37	-0.43	-0.47	-0.40	-0.25	-0.25	-0.22	-0.27	-0.38	-0.49
FSM 22	-0.28	-0.31	-0.32	-0.24	-0.20	-0.16	-0.14	-0.11	-0.25	-0.33
FSM 23	-0.35	-0.48	-0.41	-0.38	-0.35	-0.22	-0.27	-0.20	-0.37	-0.47
FSM 24	-0.31	-0.35	-0.32	-0.36	-0.25	-0.25	-0.25	-0.14	-0.27	-0.43
FSM 25	-0.31	-0.40	-0.42	-0.29	-0.25	-0.35	-0.24	-0.24	-0.35	-0.43
FSM 26	-0.09	-0.23	-0.00	-0.09	-0.04	0.05	0.04	0.11	-0.03	-0.17
FSM 27	-0.29	-0.48	-0.33	-0.39	-0.33	-0.36	-0.30	-0.34	-0.37	-0.35
FSM 28	-0.22	-0.32	-0.26	0.04	-0.20	-0.28	-0.15	-0.15	-0.20	-0.20
FSM 29	-0.26	-0.43	-0.30	-0.34	-0.27	-0.20	-0.27	-0.19	-0.28	-0.41
FSM 30	-0.12	-0.29	-0.14	-0.18	-0.16	-0.13	-0.09	0.02	-0.16	-0.18
FSM 31	-0.16	-0.19	-0.23	-0.15	-0.16	-0.17	-0.05	-0.12	-0.14	-0.24
FSM 32	-0.11	-0.24	-0.20	-0.16	-0.11	-0.08	0.04	0.06	-0.07	-0.23
FSM 33	-0.17	-0.41	-0.25	0.07	-0.20	-0.20	-0.18	-0.16	-0.26	-0.34
FSM 34	-0.22	-0.39	-0.30	-0.25	-0.21	-0.22	-0.15	-0.14	-0.18	-0.23
FSM 35	-0.17	-0.05	-0.17	-0.21	-0.10	-0.06	-0.04	0.01	-0.16	-0.24
FSM 36	-0.37	-0.48	-0.41	-0.31	-0.28	-0.30	-0.27	-0.17	-0.31	-0.48
FSM 37	-0.26	-0.42	-0.26	0.10	-0.19	-0.22	-0.17	-0.13	-0.25	-0.34
FSM 38	-0.48	-0.49	-0.41	-0.38	-0.34	-0.31	-0.38	-0.33	-0.36	-0.53
FSM 39	-0.37	-0.53	-0.39	-0.32	-0.25	-0.26	-0.18	-0.14	-0.34	-0.44
FSM 40	-0.33	-0.46	-0.33	-0.35	-0.21	-0.26	-0.19	-0.17	-0.30	-0.40
FSM 41	0.09	-0.10	-0.01	-0.04	-0.01	-0.00	0.21	0.07	0.08	-0.00
FSM 42	-0.25	-0.40	-0.34	-0.27	-0.21	-0.30	-0.24	-0.16	-0.30	-0.32
FSM 43	-0.27	-0.43	-0.38	-0.32	-0.24	-0.26	-0.27	-0.18	-0.34	-0.39
FSM 44	-0.38	0.00	-0.41	-0.24	-0.23	-0.20	-0.29	-0.15	-0.29	-0.43
FSM 45	-0.34	-0.45	-0.41	-0.37	-0.28	-0.17	-0.27	-0.27	-0.27	-0.48
FSM 46	-0.28	-0.35	-0.31	-0.22	-0.20	-0.15	-0.20	-0.14	-0.23	-0.44
FSM 47	-0.47	-0.42	-0.49	-0.43	-0.30	-0.29	-0.36	-0.30	-0.39	-0.53
FSM 48	-0.41	-0.41	-0.48	-0.34	-0.02	0.01	-0.23	-0.01	-0.38	-0.45
FSM 49	-0.33	-0.47	-0.39	-0.45	-0.27	-0.29	-0.24	-0.20	-0.35	-0.39
FSM 50	-0.03	-0.07	-0.21	-0.25	-0.16	-0.19	-0.08	-0.15	-0.09	-0.16

Appendix I.1 Contrast Model Matrix for Data Set 1, 1 of 5.

	List 11	List 12	List 13	List 14	List 15	List 16	List 17	List 18	List 19	List 20
FSM 1	-0.17	-0.05	-0.28	-0.02	-0.13	-0.00	0.03	-0.22	-0.07	-0.25
FSM 2	-0.10	-0.30	-0.57	-0.28	-0.16	-0.25	-0.32	-0.50	-0.26	-0.35
FSM 3	-0.23	-0.17	-0.38	-0.09	-0.10	-0.12	-0.17	-0.28	-0.19	-0.29
FSM 4	-0.24	-0.19	-0.07	-0.12	-0.18	-0.15	-0.25	-0.37	-0.22	-0.30
FSM 5	-0.26	-0.22	-0.41	-0.16	-0.06	-0.07	-0.28	-0.40	-0.27	-0.11
FSM 6	-0.35	-0.26	-0.54	-0.25	-0.19	0.11	-0.43	-0.48	-0.35	-0.15
FSM 7	-0.15	-0.07	-0.20	0.00	-0.05	0.10	0.04	-0.23	-0.04	-0.02
FSM 8	-0.19	-0.07	-0.37	-0.03	-0.11	0.03	-0.29	-0.37	-0.28	-0.03
FSM 9	-0.07	-0.19	-0.38	-0.11	-0.15	-0.11	-0.21	-0.32	-0.19	0.01
FSM 10	-0.09	-0.12	-0.26	0.01	-0.13	0.05	-0.05	-0.18	-0.06	-0.00
FSM 11	1.00	-0.06	-0.46	-0.11	-0.12	-0.16	-0.30	-0.37	-0.24	0.10
FSM 12	0.00	1.00	-0.39	-0.06	-0.10	-0.02	-0.22	-0.28	-0.20	-0.14
FSM 13	-0.36	-0.31	1.00	-0.28	-0.27	-0.28	-0.34	-0.40	-0.34	-0.35
FSM 14	-0.17	-0.18	-0.47	1.00	-0.04	-0.13	-0.19	-0.39	-0.30	-0.32
FSM 15	-0.33	-0.29	-0.53	-0.15	1.00	-0.26	-0.32	-0.42	-0.33	-0.45
FSM 16	-0.12	-0.06	-0.37	-0.02	-0.11	1.00	-0.16	-0.25	-0.06	-0.14
FSM 17	-0.14	-0.14	-0.35	0.11	-0.01	-0.03	1.00	-0.05	-0.06	-0.17
FSM 18	-0.29	-0.25	-0.44	-0.18	-0.20	-0.20	-0.15	1.00	-0.35	-0.31
FSM 19	-0.18	-0.17	-0.40	-0.15	-0.10	-0.03	-0.17	-0.36	1.00	-0.31
FSM 20	-0.00	-0.24	-0.47	-0.28	-0.27	-0.19	-0.35	-0.42	-0.35	1.00
FSM 21	-0.38	-0.30	-0.56	-0.25	-0.15	-0.28	-0.37	-0.45	-0.40	-0.35
FSM 22	-0.21	-0.11	-0.51	-0.10	-0.25	-0.03	-0.25	-0.39	-0.25	-0.29
FSM 23	-0.34	-0.30	-0.55	-0.28	-0.26	-0.25	-0.36	-0.46	-0.36	-0.34
FSM 24	-0.24	-0.24	-0.53	-0.11	-0.17	-0.22	-0.30	-0.41	-0.27	-0.39
FSM 25	-0.28	-0.29	-0.49	-0.18	-0.19	-0.28	-0.28	-0.49	-0.32	-0.42
FSM 26	-0.08	-0.06	-0.37	0.02	-0.11	0.11	-0.12	-0.24	-0.15	-0.05
FSM 27	-0.20	-0.31	-0.43	-0.32	-0.31	-0.25	-0.24	-0.15	-0.34	-0.25
FSM 28	-0.25	-0.27	-0.12	-0.12	-0.11	-0.17	-0.16	-0.32	-0.17	-0.36
FSM 29	-0.33	-0.31	-0.54	-0.09	-0.15	-0.25	-0.36	-0.40	-0.25	-0.43
FSM 30	-0.09	-0.03	-0.32	-0.01	-0.10	-0.01	-0.08	-0.11	-0.15	-0.26
FSM 31	-0.18	-0.17	-0.40	-0.15	-0.10	-0.03	-0.17	-0.36	1.00	-0.31
FSM 32	-0.19	-0.03	-0.38	-0.09	-0.15	0.13	-0.19	-0.27	-0.05	-0.27
FSM 33	-0.20	-0.15	0.08	-0.18	-0.17	-0.19	-0.18	-0.35	-0.15	-0.31
FSM 34	-0.22	-0.24	-0.36	-0.13	-0.11	-0.16	0.01	-0.28	-0.23	-0.29
FSM 35	-0.02	-0.13	-0.30	-0.01	-0.17	0.06	-0.18	-0.30	-0.12	-0.07
FSM 36	-0.32	-0.29	-0.52	-0.12	-0.14	-0.26	-0.42	-0.49	-0.39	-0.51
FSM 37	-0.23	-0.23	-0.06	-0.08	-0.11	-0.18	-0.22	-0.30	-0.23	-0.35
FSM 38	-0.33	-0.21	-0.59	-0.28	-0.18	-0.37	-0.39	-0.51	-0.41	-0.38
FSM 39	-0.30	-0.30	-0.50	-0.17	-0.16	-0.25	-0.41	-0.47	-0.42	-0.39
FSM 40	-0.34	-0.30	-0.45	-0.27	-0.21	-0.23	-0.16	-0.46	-0.32	-0.38
FSM 41	-0.04	-0.03	-0.19	0.02	-0.09	0.03	0.11	-0.16	-0.09	0.03
FSM 42	-0.30	-0.30	-0.44	-0.13	-0.12	-0.24	-0.24	-0.19	-0.34	-0.32
FSM 43	-0.35	-0.28	-0.52	-0.15	-0.16	-0.25	-0.30	-0.43	-0.36	-0.46
FSM 44	-0.03	-0.23	-0.52	-0.23	-0.29	-0.19	-0.28	-0.42	-0.30	-0.27
FSM 45	-0.38	-0.33	-0.56	-0.29	-0.22	-0.27	-0.33	-0.50	-0.33	-0.34
FSM 46	-0.27	-0.22	-0.47	-0.13	-0.07	-0.22	-0.34	-0.49	-0.29	-0.35
FSM 47	-0.37	-0.34	-0.57	-0.34	-0.26	-0.29	-0.47	-0.54	-0.34	-0.46
FSM 48	-0.41	-0.32	-0.50	-0.24	-0.14	-0.13	-0.38	-0.45	-0.40	-0.25
FSM 49	-0.38	-0.29	-0.52	-0.18	-0.17	-0.25	-0.30	-0.39	-0.39	-0.44
FSM 50	0.12	-0.15	-0.43	-0.04	-0.15	-0.09	-0.21	-0.30	-0.07	-0.12

Appendix I.2 Contrast Model Matrix for Data Set 1, 2 of 5

	List 21	List 22	List 23	List 24	List 25	List 26	List 27	List 28	List 29	List 30
FSM 1	-0.14	-0.10	-0.19	-0.14	-0.18	-0.11	-0.19	-0.19	-0.08	-0.08
FSM 2	-0.32	-0.27	-0.44	-0.29	-0.40	-0.39	-0.51	-0.46	-0.42	-0.37
FSM 3	-0.26	-0.18	-0.30	-0.16	-0.34	-0.09	-0.27	-0.27	-0.19	-0.15
FSM 4	-0.27	-0.18	-0.33	-0.30	-0.24	-0.27	-0.40	-0.13	-0.29	-0.25
FSM 5	-0.17	-0.20	-0.37	-0.22	-0.27	-0.23	-0.45	-0.36	-0.30	-0.26
FSM 6	-0.28	-0.25	-0.32	-0.28	-0.44	-0.30	-0.57	-0.50	-0.32	-0.29
FSM 7	0 03	-0.02	-0.17	-0.14	-0.16	-0.05	-0.29	-0.19	-0.16	-0.13
FSM 8	-0.20	-0.13	-0.24	-0.14	-0.26	-0.15	-0.52	-0.37	-0.23	-0.15
FSM 9	-0.22	-0.19	-0.33	-0.17	-0.32	-0.18	-0.41	-0.28	-0.24	-0.22
FSM 10	-0.15	-0.04	-0.25	-0.17	-0.20	-0.08	-0.15	-0.04	-0.19	-0.07
FSM 11	-0.24	-0.19	-0.32	-0.19	-0.28	-0.24	-0.24	-0.37	-0.29	-0.20
FSM 12	-0.13	0.02	-0.26	-0.14	-0.23	-0.18	-0.33	-0.37	-0.24	-0.09
FSM 13	-0.35	-0.37	-0.41	-0.35	-0.35	-0.38	-0.37	-0.19	-0.41	-0.30
FSM 14	-0.19	-0.16	-0.36	-0.14	-0.23	-0.23	-0.52	-0.32	-0.13	-0.17
FSM 15	-0.21	-0.42	-0.41	-0.27	-0.35	-0.41	-0.56	-0.40	-0.29	-0.31
FSM 16	-0.11	0.08	-0.22	-0.17	-0.25	-0.04	-0.26	-0.27	-0.22	-0.10
FSM 17	-0.14	-0.06	-0.20	-0.12	-0.08	-0.12	-0.08	-0.14	-0.17	-0.04
FSM 18	-0.32	-0.29	-0.34	-0.27	-0.43	-0.31	-0.09	-0.35	-0.28	-0.13
FSM 19	-0.26	-0.15	-0.28	-0.15	-0.25	-0.23	-0.32	-0.22	-0.14	-0.18
FSM 20	-0.30	-0.26	-0.35	-0.32	-0.41	-0.27	-0.36	-0.48	-0.40	-0.34
FSM 21	1.00	-0.39	-0.43	-0.33	-0.43	-0.36	-0.47	-0.50	-0.45	-0.38
FSM 22	-0.33	1.00	-0.28	-0.26	-0.26	-0.21	-0.38	-0.36	-0.30	-0.21
FSM 23	-0.36	-0.26	1.00	-0.30	-0.39	-0.31	-0.51	-0.46	-0.34	-0.33
FSM 24	-0.25	-0.29	-0.33	1 00	-0.38	-0.35	-0.45	-0.43	-0.31	-0.27
FSM 25	-0.33	-0.26	-0.40	-0.34	1.00	-0.36	-0.46	-0.44	-0.37	-0.27
FSM 26	-0.06	0.04	-0.11	-0.19	-0.18	1.00	-0.25	-0.23	-0.16	-0.07
FSM 27	-0.28	-0.31	-0.38	-0.40	-0.37	-0.34	1.00	-0.37	-0.38	-0.29
FSM 28	-0.31	-0.21	-0.31	-0.26	-0.32	-0.27	-0.28	1.00	-0.28	-0.17
FSM 29	-0.40	-0.28	-0.35	-0.26	-0.40	-0.35	-0.46	-0.40	1.00	-0.22
FSM 30	-0.19	-0.08	-0.26	-0.15	-0.13	-0.13	-0.22	-0.19	-0.11	1 00
FSM 31	-0.26	-0.15	-0.28	-0.15	-0.25	-0.23	-0.32	-0.22	-0.14	-0.18
FSM 32	-0.07	-0.05	-0.27	-0.21	-0.24	-0.08	-0.26	-0.19	-0.22	-0.10
FSM 33	-0.28	-0.27	-0.30	-0.27	-0.32	-0.33	-0.37	-0.19	-0.30	-0.22
FSM 34	-0.28	-0.14	-0.20	-0.27	-0.32	-0.25	-0.36	-0.21	-0.26	-0.21
FSM 35	-0.26	-0.11	-0.25	-0.15	-0.31	-0.17	-0.35	-0.26	-0.09	-0.19
FSM 36	-0.33	-0.41	-0.48	-0.24	-0.38	-0.40	-0.61	-0.36	-0.30	-0.29
FSM 37	-0.31	-0.21	-0.29	-0.24	-0.32	-0.30	-0.40	-0.08	-0.13	-0.19
FSM 38	-0.42	-0.36	-0.39	-0.34	-0.43	-0.46	-0.56	-0.53	-0.36	-0.39
FSM 39	-0.44	-0.30	-0.36	-0.34	-0.43	-0.38	-0.62	-0.47	-0.36	-0.32
FSM 40	-0.33	-0.29	0.04	-0 27	-0.34	-0.35	-0.40	-0.34	-0.32	-0.32
FSM 41	-0 12	0.09	-0.09	-0.12	-0.10	0.01	-0.19	-0.09	-0.11	-0.04
FSM 42	-0.27	-0.31	-0.34	-0.28	-0.31	-0.34	-0.42	-0.31	-0.31	-0.27
FSM 43	-0.35	-0.29	-0.39	-0.28	-0.40	-0.37	-0.40	-0.36	-0.29	-0.27
FSM 44	-0.30	-0.25	-0.42	-0.31	-0.41	-0.35	-0.49	-0.42	-0.35	-0.26
FSM 45	-0.31	-0.35	-0.36	-0.34	-0.25	-0.46	-0.54	-0.46	-0.39	-0.35
FSM 46	-0.15	-0.40	-0.34	-0.24	-0.34	-0.31	-0.60	-0.41	-0.22	-0.24
FSM 47	-0.36	-0.41	-0.51	-0.30	-0.46	-0.50	-0.50	-0.48	-0.40	-0.38
FSM 48	-0.38	-0.39	-0.45	-0.31	-0.47	-0.38	-0.57	-0.49	-0.35	-0.38
FSM 49	-0.31	-0.24	-0.46	-0.36	-0.38	-0.35	-0.46	-0.34	-0.34	-0.24
FSM 50	-0.18	-0.21	-0.23	-0.09	-0.32	-0.15	-0.19	-0.18	-0.21	-0.13

Appendix I 3 Contrast Model Matrix for Data Set 1, 3 of 5.

	List 31	List 32	List 33	List 34	List 35	List 36	List 37	List 38	List 39	List 40
FSM 1	-0.07	-0.01	-0.09	-0.23	-0.17	-0.10	-0.19	-0.30	-0.27	-0.24
FSM 2	-0.26	-0.31	-0.47	-0.48	-0.26	-0.35	-0.46	-0.46	-0.55	-0.46
FSM 3	-0.19	-0.15	-0.24	-0.33	-0.24	-0.14	-0.20	-0.21	-0.30	-0.28
FSM 4	-0.22	-0.20	-0.08	-0.37	-0.34	-0.14	0.02	-0.29	-0.33	-0.32
FSM 5	-0.27	-0.19	-0.33	-0.38	-0.32	-0.18	-0.27	-0.35	-0.32	-0.22
FSM 6	-0.35	-0.26	-0.41	-0.43	-0.40	-0.31	-0.37	-0.41	-0.43	-0.39
FSM 7	-0.04	0.07	-0.21	-0.24	-0.15	-0.06	-0.15	-0.28	-0.07	-0.14
FSM 8	-0.28	-0.08	-0.34	-0.33	-0.28	-0.10	-0.24	-0.36	-0.24	-0.24
FSM 9	-0.19	-0.10	-0.32	-0.29	-0.27	-0.11	-0.27	-0.25	-0.33	-0.30
FSM 10	-0.06	-0.03	-0.20	-0.14	-0.16	-0.11	-0.14	-0.30	-0.21	-0.23
FSM 11	-0.24	-0.22	-0.27	-0.33	-0.22	-0.13	-0.24	-0.23	-0.30	-0.33
FSM 12	-0.20	-0.03	-0.18	-0.31	-0.24	-0.08	-0.24	-0.03	-0.29	-0.28
FSM 13	-0.34	-0.31	0.01	-0.40	-0.37	-0.26	-0.06	-0.43	-0.41	-0.35
FSM 14	-0.30	-0.21	-0.37	-0.34	-0.27	-0.03	-0.17	-0.29	-0.26	-0.38
FSM 15	-0.33	-0.35	-0.40	-0.40	-0.47	-0.18	-0.31	-0.32	-0.39	-0.37
FSM 16	-0.06	0.11	-0.25	-0.26	-0.09	-0.07	-0.21	-0.33	-0.24	-0.23
FSM 17	-0.06	-0.09	-0.11	0.08	-0.18	-0.14	-0.13	-0.14	-0.26	0.08
FSM 18	-0.35	-0.24	-0.36	-0.34	-0.36	-0.24	-0.29	-0.39	-0.35	-0.39
FSM 19	1.00	-0.02	-0.16	-0.30	-0.23	-0.16	-0.22	-0.32	-0.38	-0.28
FSM 20	-0.35	-0.30	-0.41	-0.41	-0.32	-0.39	-0.38	-0.36	-0.42	-0.37
FSM 21	-0.40	-0.29	-0.46	-0.47	-0.49	-0.32	-0.46	-0.47	-0.56	-0.44
FSM 22	-0.25	-0.15	-0.37	-0.31	-0.31	-0.24	-0.27	-0.27	-0.31	-0.34
FSM 23	-0.36	-0.32	-0.43	-0.35	-0.42	-0.35	-0.36	-0.37	-0.39	-0.01
FSM 24	-0.27	-0.31	-0.40	-0.42	-0.36	-0.16	-0.35	-0.35	-0.43	-0.36
FSM 25	-0.32	-0.30	-0.43	-0.44	-0.44	-0.26	-0.38	-0.42	-0.50	-0.35
FSM 26	-0.15	0.05	-0.28	-0.23	-0.17	-0.09	-0.19	-0.26	-0.25	-0.24
FSM 27	-0.34	-0.24	-0.41	-0.40	-0.43	-0.34	-0.42	-0.43	-0.47	-0.36
FSM 28	-0.17	-0.14	-0.14	-0.24	-0.29	-0.06	0.01	-0.37	-0.39	-0.23
FSM 29	-0.25	-0.27	-0.40	-0.40	-0.34	-0.18	-0.21	-0.34	-0.42	-0.33
FSM 30	-0.15	-0.05	-0.22	-0.25	-0.24	-0.07	-0.15	-0.24	-0.30	-0.28
FSM 31	1.00	-0.02	-0.16	-0.30	-0.23	-0.16	-0.22	-0.32	-0.38	-0.28
FSM 32	-0.05	1.00	-0.33	-0.28	-0.29	0.03	-0.22	-0.39	-0.29	-0.27
FSM 33	-0.15	-0.27	1.00	-0.36	-0.30	-0.19	-0.03	-0.37	-0.38	-0.31
FSM 34	-0.23	-0.18	-0.32	1.00	-0.33	-0.06	-0.19	-0.35	-0.30	0.01
FSM 35	-0.12	-0.18	-0.22	-0.28	1.00	-0.17	-0.13	-0.25	-0.32	-0.29
FSM 36	-0.39	-0.23	-0.44	-0.37	-0.50	1.00	-0.30	-0.45	-0.39	-0.38
FSM 37	-0.23	-0.19	-0.08	-0.29	-0.28	-0.11	1.00	-0.24	-0.18	-0.29
FSM 38	-0.41	-0.47	-0.49	-0.52	-0.44	-0.38	-0.34	1.00	-0.44	-0.44
FSM 39	-0.42	-0.31	-0.47	-0.43	-0.47	-0.17	-0.21	-0.38	1.00	-0.43
FSM 40	-0.32	-0.28	-0.39	-0.19	-0.42	-0.23	-0.33	-0.39	-0.45	1.00
FSM 41	-0.09	-0.06	-0.15	0.07	-0.06	-0.11	-0.05	-0.13	-0.17	-0.12
FSM 42	-0.34	-0.29	-0.36	-0.32	-0.36	0.01	-0.26	-0.40	-0.40	-0.34
FSM 43	-0.36	-0.25	-0.37	-0.34	-0.37	-0.06	-0.29	-0.42	-0.44	-0.38
FSM 44	-0.30	-0.26	-0.41	-0.38	-0.24	-0.31	-0.34	-0.42	-0.40	-0.46
FSM 45	-0.33	-0.38	-0.47	-0.42	-0.40	-0.35	-0.38	-0.32	-0.41	-0.39
FSM 46	-0.29	-0.27	-0.41	-0.40	-0.36	-0.19	-0.25	-0.25	-0.41	-0.38
FSM 47	-0.34	-0.33	-0.52	-0.53	-0.56	-0.36	-0.44	-0.47	-0.54	-0.48
FSM 48	-0.40	-0.35	-0.38	-0.48	-0.41	-0.19	-0.37	-0.48	-0.40	-0.43
FSM 49	-0.39	-0.24	-0.44	-0.36	-0.47	-0.26	-0.32	-0.55	-0.52	-0.34
FSM 50	-0.07	-0.12	-0.30	-0.27	-0.16	-0.14	-0.20	-0.24	-0.35	-0.27

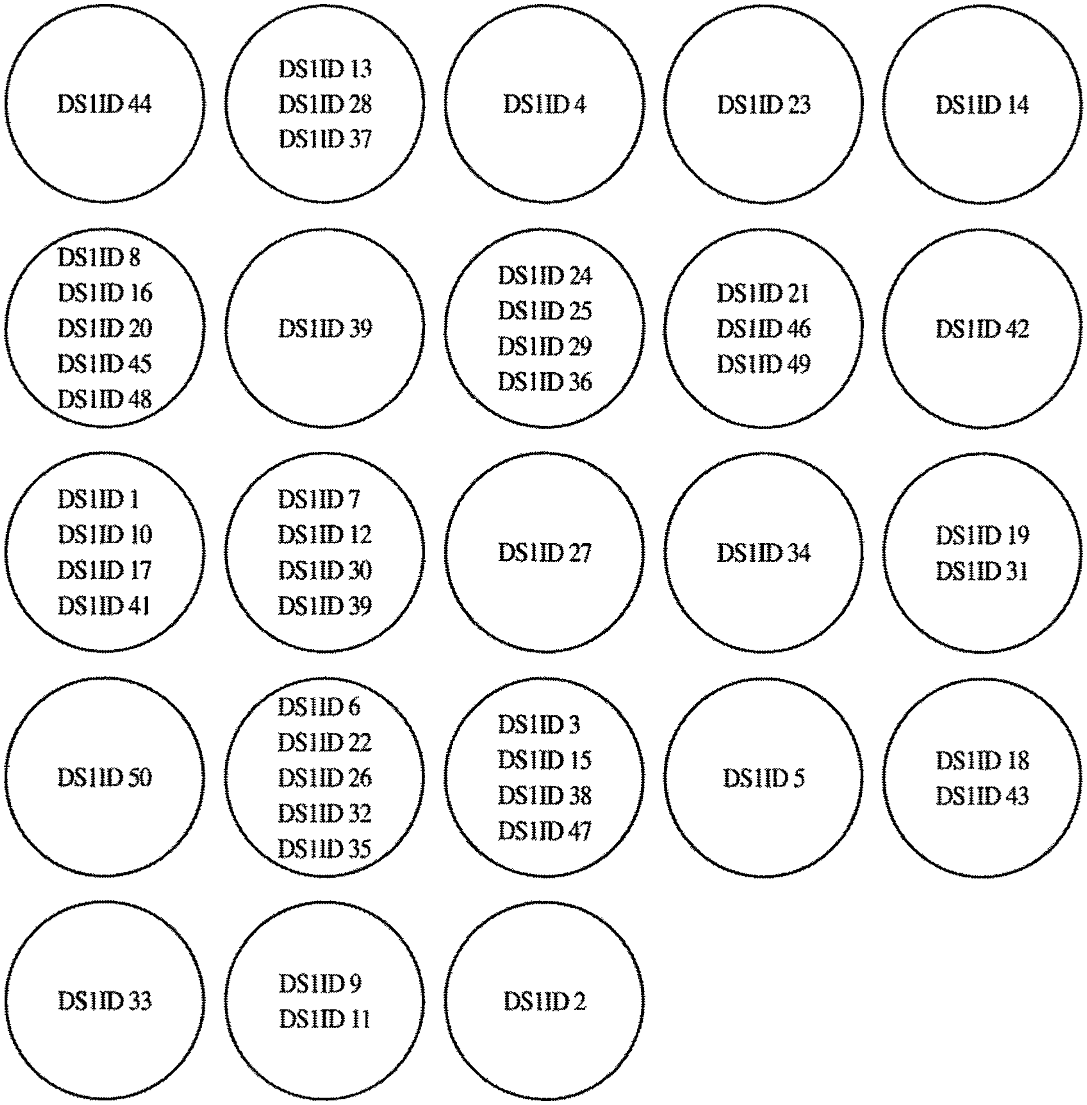
Appendix I.4 Contrast Model Matrix for Data Set 1, 4 of 5.

	List 41	List 42	List 43	List 44	List 45	List 46	List 47	List 48	List 49	List 50
FSM 1	-0.02	-0.12	-0.10	-0.21	-0.21	-0.05	-0.37	-0.17	-0.14	0.03
FSM 2	-0.35	-0.44	-0.45	0.05	-0.42	-0.24	-0.34	-0.29	-0.37	-0.22
FSM 3	-0.15	-0.26	-0.33	-0.32	-0.27	-0.08	-0.41	-0.27	-0.22	-0.21
FSM 4	-0.28	-0.27	-0.30	-0.18	-0.30	-0.11	-0.34	-0.20	-0.38	-0.30
FSM 5	-0.28	-0.25	-0.26	-0.27	-0.30	-0.12	-0.30	0.12	-0.24	-0.26
FSM 6	-0.36	-0.44	-0.39	-0.33	-0.29	-0.16	-0.35	0.03	-0.32	-0.34
FSM 7	0.01	-0.19	-0.19	-0.22	-0.18	0.01	-0.26	0.04	-0.12	-0.11
FSM 8	-0.24	-0.23	-0.25	-0.23	-0.33	-0.08	-0.31	0.09	-0.19	-0.27
FSM 9	-0.17	-0.28	-0.34	-0.22	-0.19	-0.06	-0.31	-0.23	-0.27	-0.15
FSM 10	-0.03	-0.07	-0.16	-0.20	-0.26	-0.14	-0.31	-0.04	-0.10	-0.01
FSM 11	-0.27	-0.32	-0.36	0.06	-0.34	-0.13	-0.31	-0.31	-0.32	-0.03
FSM 12	-0.20	-0.27	-0.24	-0.15	-0.27	-0.05	-0.24	-0.16	-0.15	-0.18
FSM 13	-0.34	-0.35	-0.43	-0.40	-0.44	-0.26	-0.44	-0.30	-0.35	-0.38
FSM 14	-0.29	-0.23	-0.22	-0.26	-0.34	-0.11	-0.38	-0.20	-0.22	-0.21
FSM 15	-0.42	-0.28	-0.31	-0.50	-0.38	-0.09	-0.42	-0.17	-0.26	-0.36
FSM 16	-0.18	-0.23	-0.25	-0.14	-0.23	-0.09	-0.22	0.05	-0.14	-0.15
FSM 17	0.00	-0.09	-0.13	-0.06	-0.12	-0.06	-0.33	-0.12	-0.05	-0.15
FSM 18	-0.30	-0.06	-0.37	-0.36	-0.41	-0.23	-0.38	-0.25	-0.20	-0.30
FSM 19	-0.24	-0.28	-0.33	-0.23	-0.22	-0.08	-0.22	-0.23	-0.28	-0.10
FSM 20	-0.28	-0.36	-0.49	-0.27	-0.35	-0.23	-0.39	-0.10	-0.35	-0.23
FSM 21	-0.40	-0.39	-0.45	-0.43	-0.38	-0.17	-0.41	-0.37	-0.31	-0.36
FSM 22	-0.22	-0.33	-0.33	-0.28	-0.35	-0.28	-0.38	-0.29	-0.18	-0.30
FSM 23	-0.32	-0.38	-0.44	-0.46	-0.37	-0.24	-0.51	-0.36	-0.35	-0.32
FSM 24	-0.33	-0.35	-0.34	-0.38	-0.38	-0.20	-0.34	-0.29	-0.37	-0.24
FSM 25	-0.35	-0.32	-0.40	-0.45	-0.24	-0.22	-0.47	-0.39	-0.33	-0.37
FSM 26	-0.07	-0.23	-0.22	-0.14	-0.28	-0.03	-0.34	-0.06	-0.13	-0.09
FSM 27	-0.34	-0.36	-0.34	-0.43	-0.41	-0.35	-0.40	-0.39	-0.36	-0.26
FSM 28	-0.23	-0.17	-0.20	-0.26	-0.34	-0.20	-0.34	-0.28	-0.12	-0.16
FSM 29	-0.35	-0.34	-0.31	-0.41	-0.38	-0.12	-0.35	-0.23	-0.31	-0.28
FSM 30	-0.15	-0.14	-0.18	-0.17	-0.25	-0.04	-0.33	-0.26	-0.04	-0.12
FSM 31	-0.24	-0.28	-0.33	-0.23	-0.22	-0.08	-0.22	-0.23	-0.28	-0.10
FSM 32	-0.23	-0.24	-0.19	-0.21	-0.34	-0.08	-0.23	-0.18	-0.11	-0.15
FSM 33	-0.28	-0.29	-0.33	-0.31	-0.36	-0.18	-0.41	-0.18	-0.30	-0.27
FSM 34	-0.08	-0.19	-0.18	-0.21	-0.28	-0.15	-0.39	-0.24	-0.14	-0.23
FSM 35	-0.12	-0.24	-0.21	0.03	-0.20	-0.05	-0.40	-0.06	-0.24	-0.12
FSM 36	-0.45	-0.16	-0.21	-0.49	-0.47	-0.20	-0.48	-0.23	-0.35	-0.34
FSM 37	-0.24	-0.24	-0.23	-0.31	-0.32	-0.09	-0.39	-0.24	-0.18	-0.24
FSM 38	-0.38	-0.44	-0.49	-0.49	-0.33	-0.18	-0.49	-0.39	-0.52	-0.37
FSM 39	-0.37	-0.40	-0.40	-0.41	-0.35	-0.25	-0.47	-0.24	-0.44	-0.37
FSM 40	-0.32	-0.36	-0.39	-0.47	-0.36	-0.22	-0.46	-0.32	-0.23	-0.32
FSM 41	1.00	-0.10	-0.18	-0.15	-0.09	-0.00	-0.28	-0.05	-0.11	-0.00
FSM 42	-0.28	1.00	0.07	-0.34	-0.38	-0.19	-0.42	-0.21	-0.19	-0.28
FSM 43	-0.36	0.02	1.00	-0.38	-0.41	-0.22	-0.34	-0.22	-0.15	-0.31
FSM 44	-0.38	-0.36	-0.37	1.00	-0.36	-0.27	-0.35	-0.28	-0.35	-0.20
FSM 45	-0.35	-0.42	-0.41	-0.42	1.00	-0.18	-0.39	-0.35	-0.45	-0.41
FSM 46	-0.33	-0.33	-0.33	-0.41	-0.28	1.00	-0.32	-0.28	-0.33	-0.26
FSM 47	-0.50	-0.47	-0.43	-0.43	-0.40	-0.24	1.00	-0.39	-0.47	-0.41
FSM 48	-0.42	-0.35	-0.35	-0.39	-0.46	-0.27	-0.44	1.00	-0.35	-0.43
FSM 49	-0.39	-0.28	-0.20	-0.45	-0.51	-0.27	-0.51	-0.30	1.00	-0.35
FSM 50	-0.14	-0.24	-0.22	-0.01	-0.34	-0.05	-0.37	-0.31	-0.23	1.00

Appendix I.5 Contrast Model Matrix for Data Set 1, 5 of 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1							■																																										
2																																												■					
3																																														■			
4																																					■												
5																																																■	
6																■																																	
7								■																																									
8						■																																											
9																				■																													
10							■																																										
11																				■																													
12								■																																									
13				■																																													
14																																				■													
15																																														■			
16						■																																											
17							■																																										
18																																										■							
19																																■																	
20						■																																											
21															■																																		
22																■																																	
23																																								■									
24														■																																			
25														■																																			
26																■																																	
27																		■																															
28				■																																													
29														■																																			
30								■																																									
31																																■																	
32																■																																	
33													■																																				
34																	■																																
35																■																																	
36														■																																			
37				■																																													
38																																														■			
39								■																																									
40																							■																										
41							■																																										
42																																											■						
43																																										■							
44		■																																															
45						■																																											
46															■																																		
47																																														■			
48						■																																											
49															■																																		
50												■																																					

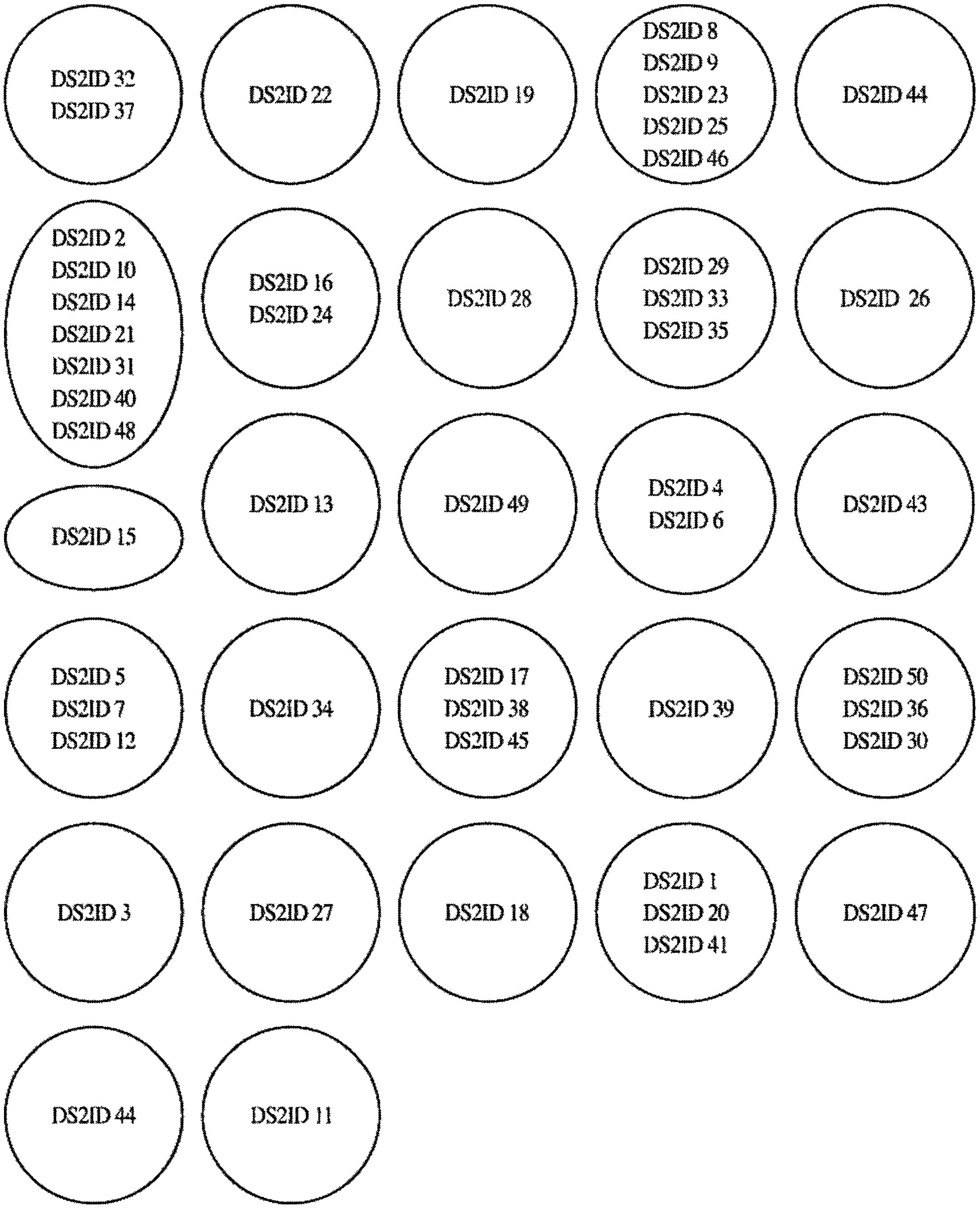
Appendix I.6 Contrast Model Similarity Grid, Data Set 1.



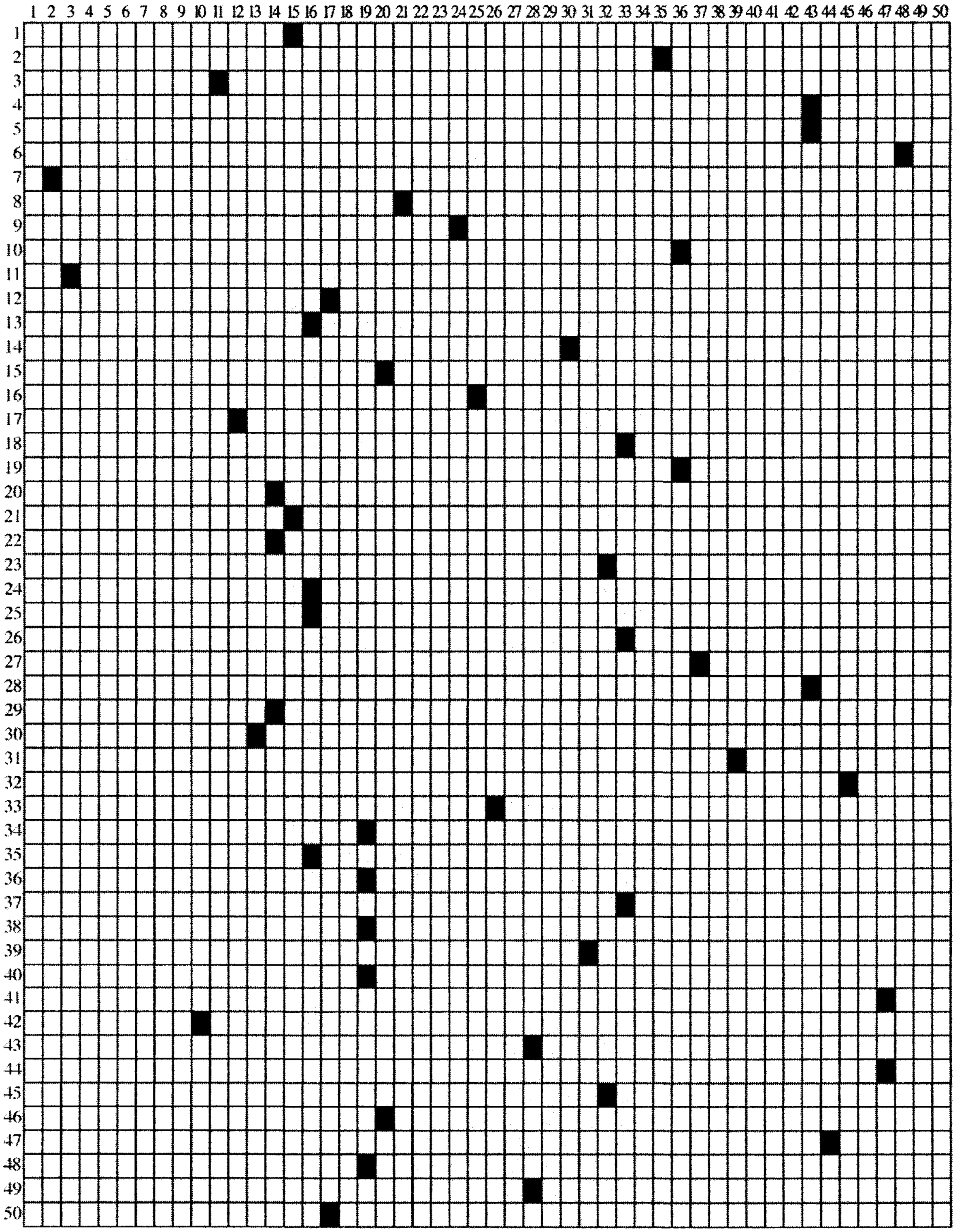
Appendix I 7 Contrast Model Similar Document Clusters, Data Set 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																						■													
2											■																																							
3																																	■																	
4																						■																												
5																								■																										
6																						■																												
7																								■																										
8							■																																											
9							■																																											
10											■																																							
11																																																■		
12																								■																										
13																					■																													
14											■																																							
15														■																																				
16												■																																						
17																														■																				
18																																			■															
19						■																																												
20																																					■													
21											■																																							
22				■																																														
23							■																																											
24												■																																						
25							■																																											
26																			■																															
27																																		■																
28													■																																					
29																		■																																
30																															■																			
31											■																																							
32	■																																																	
33																		■																																
34																											■																							
35																		■																																
36																															■																			
37	■																																																	
38																														■																				
39																																■																		
40											■																																							
41																																					■													
42																																																		
43																							■																											
44									■																																						■			
45																														■																				
46							■																																											
47																																											■							
48											■																																							
49																	■																																	
50																															■																			

Appendix I.8 Contrast Model Similarity Grid, Data Set 2.

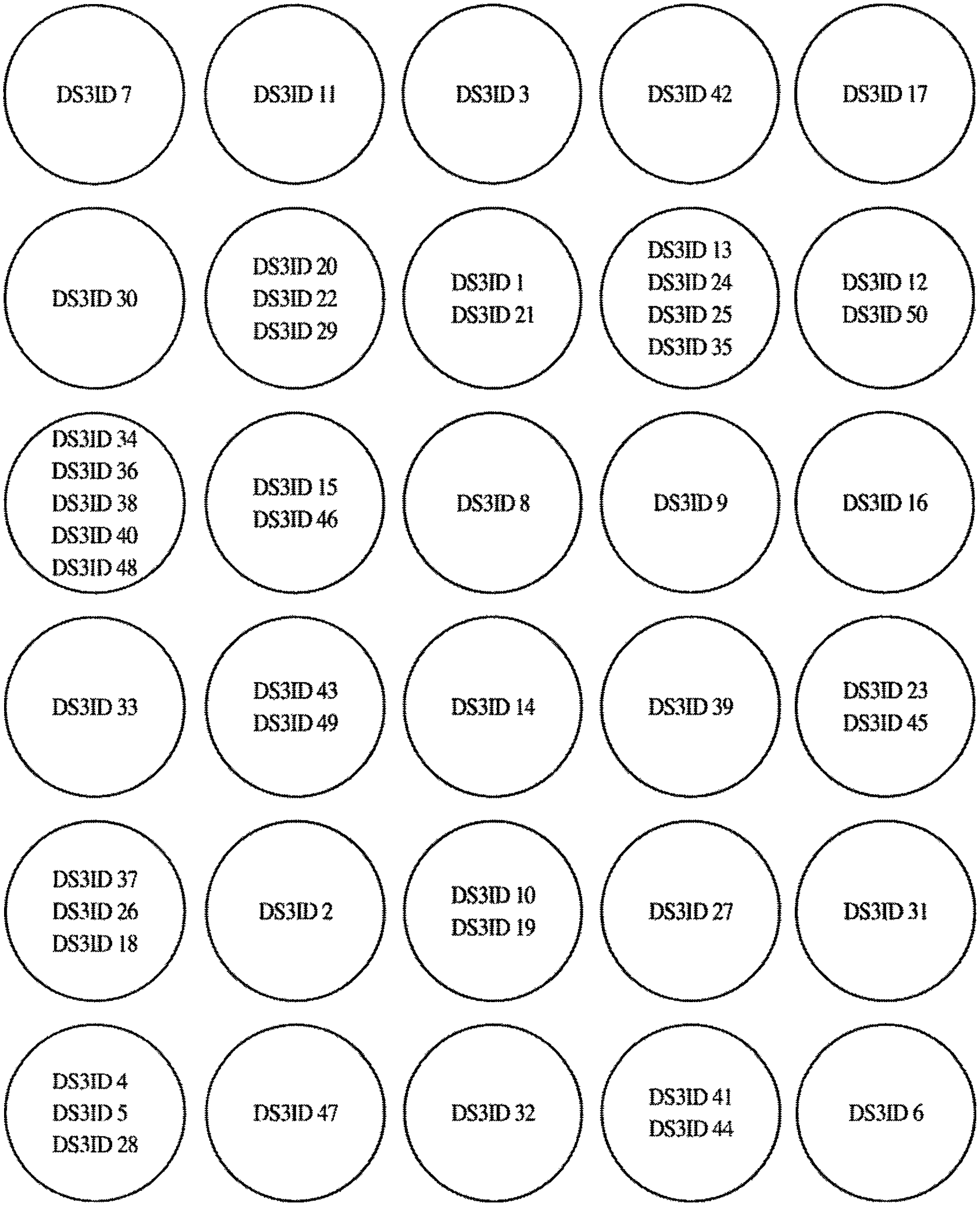


Appendix I.9 Contrast Model Similar Document Clusters, Data Set 2.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1															■																																			
2																																			■															
3											■																																							
4																																											■							
5																																											■							
6																																														■		
7		■																																																
8																					■																												
9																								■																									
10																																				■													
11			■																																															
12																	■																																
13																■																																	
14																														■																			
15																				■																													
16																									■																								
17												■																																					
18																																	■																
19																																				■													
20														■																																			
21															■																																		
22														■																																			
23																																■																	
24																■																																	
25																■																																	
26																																	■																
27																																					■												
28																																											■						
29														■																																			
30													■																																				
31																																							■										
32																																													■				
33																										■																							
34																			■																														
35																■																																	
36																			■																														
37																																	■																
38																			■																														
39																															■																		
40																			■																														
41																																															■		
42										■																																								
43																												■																					
44																																															■		
45																																■																	
46																				■																													
47																																												■					
48																			■																														
49																												■																					
50																	■																																

Appendix I.10 Contrast Model Similarity Grid, Data Set 3.

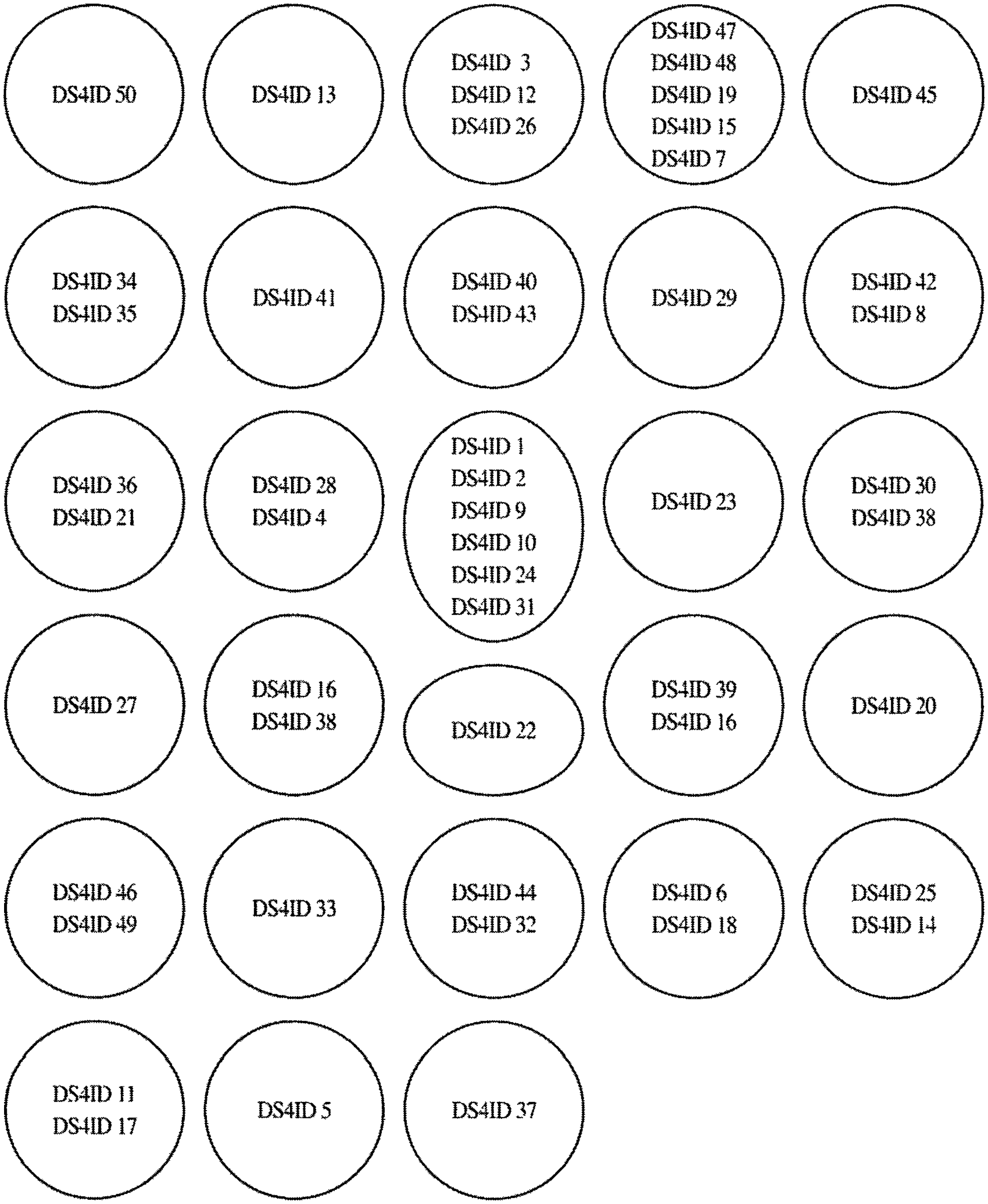


DS3ID 7	DS3ID 11	DS3ID 3	DS3ID 42	DS3ID 17
DS3ID 30	DS3ID 20 DS3ID 22 DS3ID 29	DS3ID 1 DS3ID 21	DS3ID 13 DS3ID 24 DS3ID 25 DS3ID 35	DS3ID 12 DS3ID 50
DS3ID 34 DS3ID 36 DS3ID 38 DS3ID 40 DS3ID 48	DS3ID 15 DS3ID 46	DS3ID 8	DS3ID 9	DS3ID 16
DS3ID 33	DS3ID 43 DS3ID 49	DS3ID 14	DS3ID 39	DS3ID 23 DS3ID 45
DS3ID 37 DS3ID 26 DS3ID 18	DS3ID 2	DS3ID 10 DS3ID 19	DS3ID 27	DS3ID 31
DS3ID 4 DS3ID 5 DS3ID 28	DS3ID 47	DS3ID 32	DS3ID 41 DS3ID 44	DS3ID 6

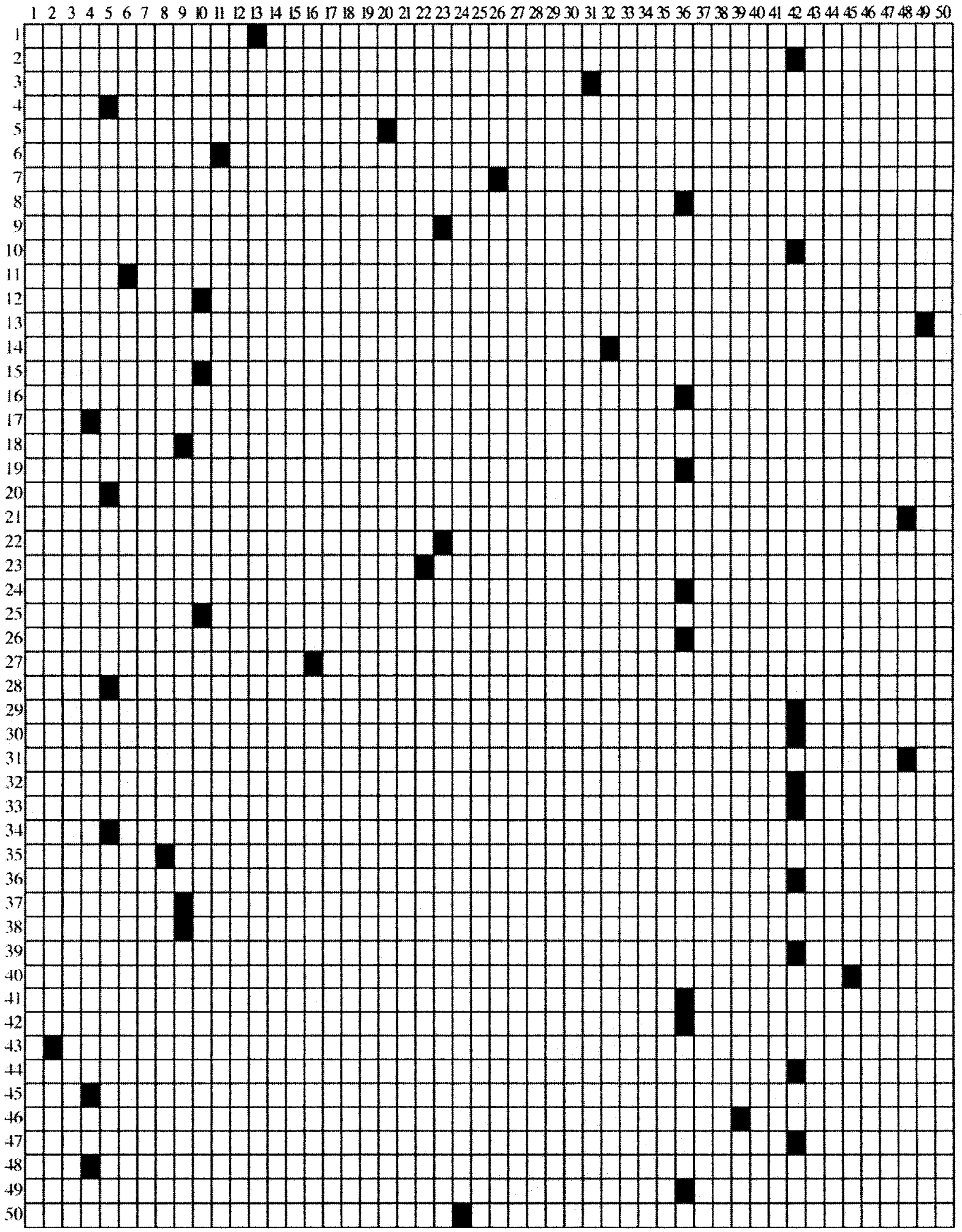
Appendix I.11 Contrast Model Similar Document Clusters, Data Set 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																							■																											
2																							■																											
3							■																																											
4																						■																												
5																																																■		
6																																										■									
7								■																																										
8																			■																															
9																							■																											
10																							■																											
11																																															■			
12							■																																											
13			■																																															
14																																												■						
15								■																																										
16																																	■																		
17																																															■			
18																																										■									
19								■																																										
20																																				■														
21																				■																														
22																												■																						
23																								■																										
24																							■																											
25																																												■						
26							■																																											
27																														■																				
28																						■																												
29															■																																			
30																											■																							
31																							■																											
32																																								■										
33																																							■											
34											■																																							
35											■																																							
36																				■																														
37																																																	■	
38																											■																							
39																																	■																		
40													■																																					
41												■																																						
42																			■																															
43													■																																					
44																																								■										
45									■																																									
46																																					■													
47								■																																										
48								■																																										
49																																					■													
50		■																																																

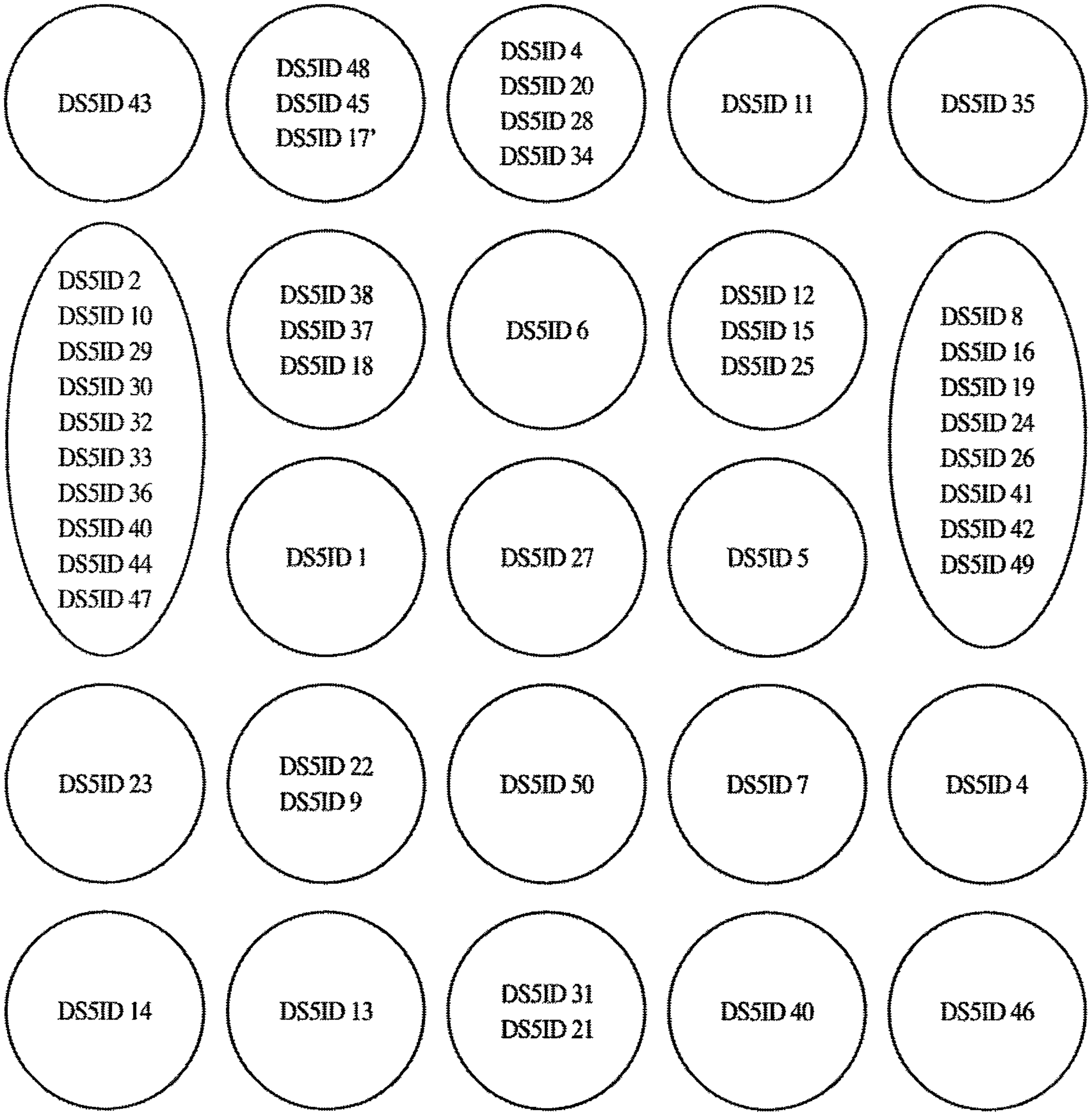
Appendix I.12 Contrast Model Similarity Grid, Data Set 4.



Appendix I.13 Contrast Model Similar Document Clusters, Data Set 4.



Appendix I.14 Contrast Model Similarity Grid, Data Set 5.



Appendix I.15 Contrast Model Similar Document Clusters, Data Set 5.

APPENDIX J

RATIO MODEL EXTENDED RESULTS

DATA SETS 1-5

	List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
FSM 1	1.00	0.16	0.30	0.17	0.20	0.15	0.48	0.23	0.29	0.42
FSM 2	0.36	1.00	0.30	0.31	0.33	0.30	0.41	0.35	0.42	0.32
FSM 3	0.41	0.19	1.00	0.21	0.26	0.21	0.42	0.29	0.33	0.40
FSM 4	0.42	0.34	0.37	1.00	0.40	0.33	0.45	0.42	0.45	0.39
FSM 5	0.40	0.28	0.35	0.32	1.00	0.41	0.49	0.49	0.38	0.38
FSM 6	0.35	0.28	0.31	0.29	0.45	1.00	0.43	0.54	0.35	0.32
FSM 7	0.52	0.21	0.33	0.22	0.31	0.24	1.00	0.37	0.38	0.44
FSM 8	0.40	0.27	0.35	0.32	0.46	0.48	0.52	1.00	0.37	0.37
FSM 9	0.43	0.35	0.37	0.32	0.33	0.27	0.50	0.34	1.00	0.39
FSM 10	0.42	0.14	0.30	0.17	0.21	0.15	0.41	0.23	0.25	1.00
FSM 11	0.38	0.37	0.34	0.27	0.32	0.27	0.42	0.37	0.41	0.37
FSM 12	0.43	0.24	0.35	0.27	0.30	0.26	0.45	0.39	0.34	0.36
FSM 13	0.40	0.24	0.37	0.45	0.34	0.28	0.47	0.39	0.38	0.36
FSM 14	0.40	0.27	0.37	0.31	0.34	0.29	0.44	0.42	0.37	0.36
FSM 15	0.36	0.31	0.33	0.30	0.40	0.36	0.41	0.40	0.34	0.29
FSM 16	0.45	0.24	0.36	0.27	0.36	0.42	0.51	0.42	0.37	0.42
FSM 17	0.47	0.19	0.32	0.19	0.22	0.15	0.48	0.24	0.29	0.41
FSM 18	0.39	0.20	0.35	0.24	0.28	0.24	0.42	0.33	0.34	0.37
FSM 19	0.43	0.25	0.34	0.25	0.28	0.23	0.46	0.30	0.34	0.39
FSM 20	0.35	0.32	0.33	0.30	0.44	0.45	0.45	0.48	0.46	0.37
FSM 21	0.36	0.29	0.31	0.29	0.39	0.36	0.43	0.38	0.34	0.30
FSM 22	0.40	0.31	0.36	0.33	0.37	0.34	0.47	0.42	0.38	0.37
FSM 23	0.38	0.24	0.33	0.28	0.31	0.34	0.41	0.39	0.33	0.32
FSM 24	0.39	0.29	0.37	0.28	0.35	0.32	0.42	0.41	0.37	0.33
FSM 25	0.40	0.28	0.33	0.31	0.36	0.28	0.43	0.37	0.34	0.33
FSM 26	0.42	0.18	0.38	0.20	0.25	0.21	0.44	0.30	0.32	0.40
FSM 27	0.41	0.23	0.36	0.26	0.31	0.26	0.40	0.32	0.32	0.37
FSM 28	0.41	0.22	0.34	0.33	0.28	0.20	0.42	0.30	0.33	0.42
FSM 29	0.41	0.25	0.37	0.29	0.34	0.33	0.41	0.38	0.37	0.34
FSM 30	0.40	0.15	0.31	0.17	0.20	0.16	0.37	0.26	0.26	0.38
FSM 31	0.43	0.25	0.34	0.25	0.28	0.23	0.46	0.30	0.34	0.39
FSM 32	0.46	0.25	0.36	0.26	0.31	0.27	0.51	0.38	0.38	0.40
FSM 33	0.45	0.22	0.38	0.39	0.32	0.27	0.44	0.35	0.34	0.37
FSM 34	0.40	0.18	0.30	0.21	0.25	0.20	0.41	0.29	0.32	0.40
FSM 35	0.42	0.27	0.34	0.21	0.27	0.22	0.45	0.31	0.31	0.39
FSM 36	0.36	0.25	0.32	0.32	0.36	0.32	0.40	0.41	0.37	0.30
FSM 37	0.41	0.22	0.37	0.42	0.33	0.27	0.44	0.36	0.35	0.37
FSM 38	0.32	0.25	0.33	0.30	0.34	0.32	0.37	0.35	0.35	0.28
FSM 39	0.37	0.21	0.33	0.29	0.35	0.30	0.45	0.41	0.34	0.32
FSM 40	0.40	0.25	0.37	0.28	0.38	0.31	0.46	0.40	0.37	0.35
FSM 41	0.41	0.13	0.27	0.13	0.16	0.12	0.40	0.18	0.24	0.40
FSM 42	0.43	0.27	0.36	0.32	0.37	0.29	0.43	0.40	0.36	0.38
FSM 43	0.42	0.28	0.36	0.32	0.38	0.34	0.43	0.42	0.37	0.36
FSM 44	0.36	0.47	0.33	0.34	0.38	0.35	0.41	0.42	0.37	0.33
FSM 45	0.39	0.26	0.33	0.28	0.35	0.36	0.41	0.36	0.38	0.31
FSM 46	0.40	0.28	0.36	0.33	0.37	0.35	0.43	0.40	0.38	0.32
FSM 47	0.34	0.30	0.31	0.29	0.38	0.35	0.39	0.38	0.35	0.29
FSM 48	0.35	0.30	0.30	0.33	0.50	0.49	0.42	0.51	0.34	0.31
FSM 49	0.38	0.24	0.33	0.24	0.35	0.30	0.43	0.39	0.34	0.34
FSM 50	0.45	0.22	0.29	0.16	0.21	0.16	0.39	0.22	0.29	0.40

Appendix J.1 Ratio Model Matrix for Data Set 1, 1 of 5.

	List 11	List 12	List 13	List 14	List 15	List 16	List 17	List 18	List 19	List 20
FSM 1	0.22	0.32	0.15	0.25	0.15	0.35	0.45	0.23	0.33	0.14
FSM 2	0.44	0.37	0.19	0.35	0.34	0.40	0.38	0.26	0.40	0.27
FSM 3	0.29	0.36	0.19	0.32	0.24	0.39	0.42	0.28	0.36	0.20
FSM 4	0.39	0.43	0.42	0.44	0.35	0.46	0.43	0.33	0.42	0.31
FSM 5	0.35	0.40	0.25	0.39	0.36	0.48	0.40	0.29	0.38	0.36
FSM 6	0.33	0.39	0.22	0.38	0.36	0.56	0.33	0.27	0.36	0.39
FSM 7	0.28	0.36	0.21	0.31	0.21	0.44	0.49	0.27	0.39	0.23
FSM 8	0.38	0.46	0.26	0.44	0.34	0.51	0.39	0.30	0.37	0.39
FSM 9	0.39	0.38	0.22	0.36	0.27	0.42	0.41	0.30	0.39	0.34
FSM 10	0.22	0.27	0.13	0.23	0.13	0.33	0.39	0.21	0.30	0.17
FSM 11	1.00	0.46	0.20	0.38	0.30	0.42	0.38	0.29	0.38	0.41
FSM 12	0.39	1.00	0.19	0.35	0.25	0.44	0.40	0.29	0.37	0.25
FSM 13	0.33	0.38	1.00	0.36	0.31	0.40	0.40	0.32	0.37	0.28
FSM 14	0.38	0.41	0.21	1.00	0.35	0.44	0.42	0.29	0.35	0.26
FSM 15	0.34	0.38	0.22	0.43	1.00	0.40	0.37	0.30	0.36	0.25
FSM 16	0.34	0.42	0.20	0.36	0.25	1.00	0.42	0.30	0.43	0.25
FSM 17	0.24	0.30	0.14	0.30	0.18	0.35	1.00	0.30	0.34	0.16
FSM 18	0.30	0.36	0.20	0.34	0.26	0.39	0.45	1.00	0.32	0.23
FSM 19	0.31	0.36	0.18	0.30	0.24	0.42	0.42	0.25	1.00	0.19
FSM 20	0.50	0.41	0.25	0.37	0.32	0.44	0.37	0.31	0.37	1.00
FSM 21	0.34	0.39	0.22	0.40	0.41	0.40	0.36	0.31	0.35	0.32
FSM 22	0.39	0.46	0.21	0.43	0.30	0.50	0.41	0.31	0.40	0.30
FSM 23	0.34	0.38	0.21	0.36	0.32	0.41	0.37	0.29	0.35	0.29
FSM 24	0.38	0.40	0.21	0.44	0.35	0.42	0.40	0.30	0.39	0.26
FSM 25	0.36	0.39	0.23	0.40	0.34	0.40	0.41	0.27	0.37	0.25
FSM 26	0.28	0.34	0.15	0.29	0.17	0.42	0.40	0.24	0.32	0.20
FSM 27	0.39	0.37	0.25	0.34	0.28	0.41	0.42	0.42	0.36	0.31
FSM 28	0.29	0.33	0.29	0.33	0.25	0.38	0.44	0.28	0.38	0.19
FSM 29	0.33	0.37	0.20	0.44	0.35	0.40	0.36	0.30	0.40	0.24
FSM 30	0.26	0.34	0.15	0.26	0.16	0.35	0.40	0.27	0.30	0.14
FSM 31	0.31	0.36	0.18	0.30	0.24	0.42	0.42	0.25	1.00	0.19
FSM 32	0.31	0.43	0.20	0.34	0.24	0.51	0.41	0.30	0.43	0.22
FSM 33	0.35	0.42	0.41	0.35	0.28	0.40	0.44	0.30	0.42	0.24
FSM 34	0.28	0.32	0.19	0.30	0.23	0.36	0.50	0.27	0.34	0.19
FSM 35	0.34	0.35	0.19	0.32	0.19	0.44	0.40	0.25	0.37	0.23
FSM 36	0.35	0.38	0.23	0.44	0.39	0.40	0.33	0.27	0.34	0.23
FSM 37	0.34	0.38	0.36	0.39	0.30	0.41	0.42	0.32	0.38	0.23
FSM 38	0.35	0.43	0.20	0.37	0.37	0.36	0.35	0.27	0.34	0.29
FSM 39	0.35	0.37	0.22	0.40	0.34	0.40	0.34	0.27	0.32	0.25
FSM 40	0.34	0.38	0.25	0.36	0.33	0.42	0.45	0.28	0.38	0.27
FSM 41	0.19	0.25	0.12	0.18	0.11	0.27	0.40	0.18	0.24	0.13
FSM 42	0.35	0.38	0.25	0.42	0.36	0.41	0.43	0.41	0.36	0.29
FSM 43	0.35	0.40	0.23	0.44	0.39	0.42	0.40	0.31	0.37	0.25
FSM 44	0.49	0.42	0.23	0.39	0.31	0.44	0.40	0.31	0.39	0.33
FSM 45	0.32	0.37	0.20	0.36	0.34	0.40	0.38	0.26	0.37	0.29
FSM 46	0.35	0.40	0.23	0.41	0.38	0.41	0.36	0.26	0.37	0.27
FSM 47	0.35	0.38	0.23	0.37	0.37	0.41	0.33	0.26	0.38	0.27
FSM 48	0.32	0.38	0.25	0.41	0.41	0.47	0.35	0.30	0.35	0.36
FSM 49	0.31	0.38	0.21	0.41	0.35	0.40	0.39	0.31	0.34	0.24
FSM 50	0.34	0.31	0.13	0.27	0.16	0.33	0.36	0.22	0.35	0.18

Appendix J.2 Ratio Model Matrix for Data Set 1, 2 of 5.

	List 21	List 22	List 23	List 24	List 25	List 26	List 27	List 28	List 29	List 30
FSM 1	0.13	0.19	0.16	0.17	0.16	0.38	0.18	0.27	0.20	0.41
FSM 2	0.25	0.32	0.24	0.31	0.26	0.35	0.22	0.29	0.26	0.37
FSM 3	0.17	0.25	0.20	0.25	0.19	0.45	0.23	0.31	0.25	0.44
FSM 4	0.28	0.38	0.30	0.32	0.34	0.41	0.28	0.45	0.32	0.44
FSM 5	0.29	0.34	0.25	0.32	0.30	0.42	0.24	0.33	0.30	0.42
FSM 6	0.29	0.35	0.31	0.33	0.26	0.38	0.21	0.27	0.32	0.41
FSM 7	0.19	0.27	0.20	0.21	0.21	0.44	0.19	0.31	0.22	0.42
FSM 8	0.27	0.37	0.30	0.35	0.30	0.45	0.20	0.32	0.32	0.46
FSM 9	0.21	0.30	0.23	0.29	0.23	0.42	0.22	0.34	0.27	0.42
FSM 10	0.11	0.18	0.13	0.14	0.14	0.37	0.17	0.30	0.15	0.39
FSM 11	0.23	0.32	0.25	0.30	0.26	0.40	0.30	0.31	0.27	0.44
FSM 12	0.21	0.34	0.23	0.27	0.24	0.41	0.22	0.28	0.24	0.47
FSM 13	0.24	0.29	0.26	0.29	0.28	0.37	0.29	0.42	0.27	0.43
FSM 14	0.26	0.35	0.24	0.34	0.30	0.40	0.20	0.33	0.35	0.45
FSM 15	0.33	0.28	0.27	0.34	0.30	0.33	0.21	0.32	0.34	0.39
FSM 16	0.21	0.36	0.24	0.26	0.23	0.48	0.24	0.32	0.25	0.46
FSM 17	0.13	0.22	0.16	0.19	0.19	0.38	0.22	0.30	0.18	0.43
FSM 18	0.19	0.26	0.22	0.26	0.20	0.37	0.35	0.31	0.26	0.47
FSM 19	0.17	0.27	0.21	0.26	0.22	0.38	0.22	0.33	0.27	0.42
FSM 20	0.28	0.35	0.29	0.31	0.27	0.41	0.30	0.29	0.28	0.39
FSM 21	1.00	0.31	0.28	0.34	0.28	0.36	0.27	0.29	0.28	0.37
FSM 22	0.24	1.00	0.31	0.32	0.31	0.43	0.28	0.33	0.31	0.44
FSM 23	0.25	0.34	1.00	0.31	0.27	0.38	0.23	0.29	0.31	0.40
FSM 24	0.29	0.33	0.29	1.00	0.28	0.37	0.25	0.30	0.31	0.42
FSM 25	0.26	0.34	0.26	0.30	1.00	0.37	0.25	0.31	0.29	0.42
FSM 26	0.16	0.25	0.19	0.18	0.18	1.00	0.18	0.28	0.20	0.43
FSM 27	0.26	0.31	0.26	0.26	0.26	0.38	1.00	0.33	0.27	0.41
FSM 28	0.16	0.26	0.21	0.23	0.21	0.37	0.24	1.00	0.23	0.44
FSM 29	0.22	0.32	0.28	0.32	0.26	0.37	0.24	0.32	1.00	0.44
FSM 30	0.12	0.21	0.15	0.18	0.18	0.37	0.17	0.28	0.20	1.00
FSM 31	0.17	0.27	0.21	0.26	0.22	0.38	0.22	0.33	0.27	0.42
FSM 32	0.22	0.32	0.22	0.24	0.23	0.46	0.25	0.36	0.25	0.46
FSM 33	0.21	0.28	0.25	0.27	0.24	0.37	0.24	0.39	0.26	0.43
FSM 34	0.15	0.26	0.22	0.21	0.19	0.37	0.19	0.33	0.22	0.40
FSM 35	0.14	0.25	0.19	0.22	0.18	0.40	0.18	0.29	0.25	0.40
FSM 36	0.28	0.28	0.24	0.36	0.29	0.34	0.19	0.34	0.34	0.41
FSM 37	0.20	0.31	0.26	0.28	0.25	0.38	0.23	0.44	0.33	0.45
FSM 38	0.24	0.31	0.29	0.32	0.27	0.32	0.22	0.27	0.31	0.37
FSM 39	0.20	0.31	0.27	0.29	0.24	0.35	0.17	0.29	0.28	0.41
FSM 40	0.25	0.32	0.46	0.33	0.29	0.38	0.28	0.35	0.31	0.41
FSM 41	0.08	0.16	0.12	0.12	0.12	0.35	0.12	0.23	0.13	0.35
FSM 42	0.27	0.31	0.29	0.31	0.30	0.38	0.26	0.36	0.31	0.42
FSM 43	0.27	0.35	0.29	0.35	0.29	0.37	0.30	0.35	0.34	0.43
FSM 44	0.28	0.35	0.26	0.32	0.27	0.37	0.25	0.31	0.31	0.43
FSM 45	0.27	0.30	0.28	0.30	0.34	0.32	0.22	0.30	0.28	0.38
FSM 46	0.31	0.26	0.28	0.33	0.28	0.37	0.18	0.30	0.34	0.42
FSM 47	0.30	0.31	0.26	0.36	0.28	0.31	0.27	0.30	0.31	0.39
FSM 48	0.28	0.31	0.27	0.35	0.27	0.35	0.22	0.29	0.33	0.38
FSM 49	0.27	0.35	0.24	0.29	0.28	0.36	0.25	0.35	0.30	0.42
FSM 50	0.13	0.18	0.16	0.20	0.15	0.37	0.20	0.29	0.18	0.40

Appendix J.3 Ratio Model Matrix for Data Set 1, 3 of 5

	List 31	List 32	List 33	List 34	List 35	List 36	List 37	List 38	List 39	List 40
FSM 1	0.33	0.34	0.26	0.28	0.32	0.15	0.23	0.11	0.16	0.15
FSM 2	0.40	0.37	0.27	0.29	0.40	0.26	0.27	0.21	0.21	0.23
FSM 3	0.36	0.37	0.29	0.30	0.36	0.22	0.30	0.20	0.22	0.21
FSM 4	0.42	0.43	0.46	0.35	0.37	0.36	0.51	0.29	0.31	0.31
FSM 5	0.38	0.42	0.32	0.33	0.36	0.31	0.35	0.24	0.29	0.32
FSM 6	0.36	0.40	0.31	0.31	0.32	0.30	0.33	0.25	0.27	0.28
FSM 7	0.39	0.43	0.27	0.31	0.37	0.20	0.28	0.15	0.25	0.22
FSM 8	0.37	0.46	0.31	0.34	0.37	0.33	0.36	0.23	0.32	0.31
FSM 9	0.39	0.43	0.29	0.35	0.36	0.27	0.31	0.23	0.24	0.25
FSM 10	0.30	0.30	0.20	0.29	0.30	0.13	0.21	0.10	0.15	0.14
FSM 11	0.38	0.38	0.33	0.34	0.39	0.29	0.34	0.26	0.27	0.25
FSM 12	0.37	0.43	0.33	0.32	0.36	0.25	0.30	0.27	0.23	0.22
FSM 13	0.37	0.39	0.50	0.34	0.36	0.30	0.47	0.23	0.27	0.29
FSM 14	0.35	0.39	0.29	0.34	0.38	0.35	0.38	0.25	0.30	0.24
FSM 15	0.36	0.35	0.31	0.33	0.29	0.36	0.35	0.29	0.29	0.29
FSM 16	0.43	0.50	0.29	0.34	0.43	0.25	0.31	0.18	0.25	0.24
FSM 17	0.34	0.32	0.27	0.41	0.32	0.15	0.26	0.15	0.17	0.24
FSM 18	0.32	0.37	0.28	0.33	0.33	0.24	0.31	0.19	0.24	0.21
FSM 19	1.00	0.43	0.33	0.31	0.36	0.22	0.29	0.17	0.19	0.21
FSM 20	0.37	0.39	0.31	0.33	0.37	0.26	0.32	0.27	0.28	0.29
FSM 21	0.35	0.39	0.29	0.31	0.30	0.32	0.30	0.25	0.23	0.28
FSM 22	0.40	0.44	0.31	0.37	0.37	0.30	0.36	0.28	0.30	0.29
FSM 23	0.35	0.37	0.29	0.36	0.32	0.27	0.33	0.26	0.28	0.45
FSM 24	0.39	0.37	0.31	0.32	0.35	0.34	0.33	0.26	0.26	0.28
FSM 25	0.37	0.38	0.29	0.31	0.32	0.31	0.32	0.24	0.23	0.29
FSM 26	0.32	0.39	0.22	0.30	0.34	0.17	0.25	0.14	0.18	0.17
FSM 27	0.36	0.40	0.29	0.33	0.32	0.26	0.29	0.22	0.24	0.27
FSM 28	0.38	0.39	0.34	0.36	0.34	0.26	0.41	0.17	0.20	0.24
FSM 29	0.40	0.39	0.30	0.33	0.35	0.33	0.39	0.26	0.26	0.29
FSM 30	0.30	0.33	0.23	0.28	0.29	0.16	0.24	0.13	0.15	0.15
FSM 31	1.00	0.43	0.33	0.31	0.36	0.22	0.29	0.17	0.19	0.21
FSM 32	0.43	1.00	0.26	0.33	0.34	0.28	0.31	0.16	0.23	0.23
FSM 33	0.42	0.36	1.00	0.33	0.36	0.26	0.43	0.20	0.24	0.25
FSM 34	0.34	0.35	0.25	1.00	0.31	0.23	0.30	0.15	0.21	0.29
FSM 35	0.37	0.33	0.27	0.31	1.00	0.19	0.30	0.17	0.19	0.19
FSM 36	0.34	0.41	0.29	0.34	0.27	1.00	0.36	0.24	0.29	0.29
FSM 37	0.38	0.40	0.41	0.36	0.37	0.30	1.00	0.25	0.32	0.26
FSM 38	0.34	0.31	0.28	0.28	0.32	0.28	0.35	1.00	0.28	0.27
FSM 39	0.32	0.37	0.26	0.31	0.30	0.33	0.39	0.24	1.00	0.25
FSM 40	0.38	0.39	0.31	0.43	0.33	0.31	0.34	0.25	0.25	1.00
FSM 41	0.24	0.24	0.17	0.31	0.29	0.10	0.19	0.10	0.12	0.12
FSM 42	0.36	0.38	0.32	0.37	0.36	0.41	0.37	0.24	0.27	0.29
FSM 43	0.37	0.42	0.34	0.37	0.36	0.42	0.38	0.26	0.27	0.30
FSM 44	0.39	0.40	0.31	0.34	0.40	0.30	0.34	0.24	0.28	0.25
FSM 45	0.37	0.34	0.28	0.32	0.33	0.27	0.32	0.28	0.28	0.28
FSM 46	0.37	0.38	0.29	0.32	0.34	0.32	0.37	0.30	0.26	0.27
FSM 47	0.38	0.38	0.28	0.29	0.27	0.32	0.32	0.26	0.25	0.27
FSM 48	0.35	0.36	0.34	0.29	0.32	0.38	0.34	0.24	0.30	0.28
FSM 49	0.34	0.40	0.29	0.35	0.30	0.31	0.34	0.18	0.22	0.29
FSM 50	0.35	0.32	0.21	0.28	0.34	0.16	0.24	0.14	0.15	0.16

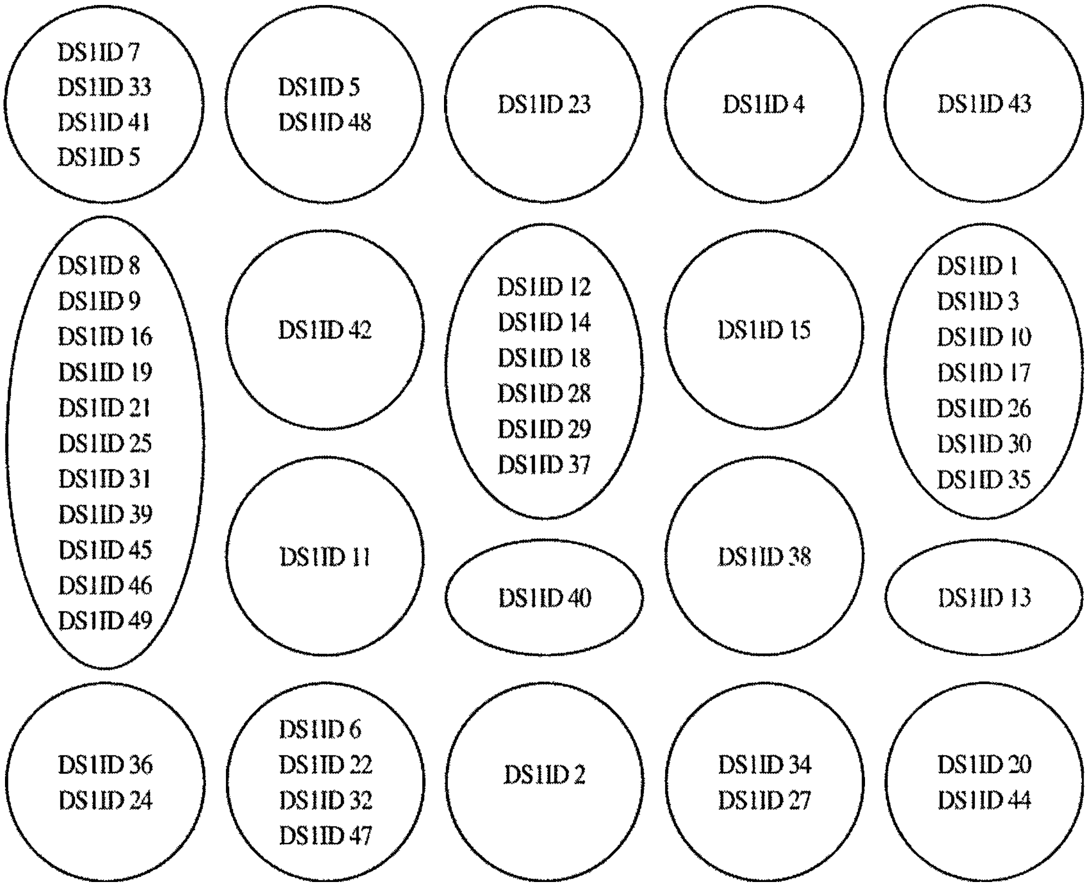
Appendix J.4 Ratio Model Matrix for Data Set 1, 4 of 5.

	List 41	List 42	List 43	List 44	List 45	List 46	List 47	List 48	List 49	List 50
FSM 1	0.49	0.19	0.15	0.15	0.15	0.21	0.08	0.12	0.17	0.44
FSM 2	0.37	0.25	0.22	0.44	0.25	0.34	0.23	0.26	0.27	0.43
FSM 3	0.46	0.23	0.17	0.19	0.21	0.29	0.12	0.16	0.23	0.40
FSM 4	0.41	0.34	0.29	0.35	0.31	0.41	0.24	0.31	0.28	0.42
FSM 5	0.40	0.32	0.28	0.29	0.28	0.38	0.23	0.40	0.31	0.42
FSM 6	0.35	0.27	0.26	0.30	0.32	0.40	0.25	0.42	0.31	0.38
FSM 7	0.52	0.21	0.17	0.18	0.20	0.27	0.13	0.19	0.22	0.42
FSM 8	0.41	0.32	0.28	0.30	0.27	0.39	0.22	0.38	0.33	0.40
FSM 9	0.45	0.26	0.20	0.26	0.28	0.34	0.18	0.21	0.25	0.45
FSM 10	0.47	0.17	0.12	0.13	0.12	0.16	0.08	0.12	0.16	0.40
FSM 11	0.41	0.26	0.21	0.39	0.24	0.34	0.20	0.21	0.25	0.51
FSM 12	0.44	0.24	0.21	0.25	0.22	0.31	0.17	0.20	0.26	0.42
FSM 13	0.40	0.29	0.23	0.25	0.24	0.34	0.20	0.26	0.29	0.39
FSM 14	0.39	0.31	0.28	0.28	0.25	0.37	0.19	0.26	0.31	0.42
FSM 15	0.33	0.34	0.30	0.23	0.28	0.43	0.23	0.34	0.34	0.37
FSM 16	0.44	0.25	0.20	0.25	0.23	0.30	0.18	0.26	0.26	0.43
FSM 17	0.51	0.21	0.16	0.19	0.18	0.21	0.09	0.14	0.20	0.37
FSM 18	0.40	0.34	0.20	0.21	0.20	0.28	0.16	0.21	0.28	0.38
FSM 19	0.41	0.22	0.17	0.22	0.22	0.29	0.17	0.17	0.21	0.45
FSM 20	0.40	0.30	0.21	0.32	0.29	0.37	0.23	0.36	0.30	0.44
FSM 21	0.35	0.31	0.26	0.28	0.31	0.42	0.25	0.28	0.34	0.38
FSM 22	0.43	0.29	0.26	0.30	0.27	0.32	0.21	0.25	0.35	0.40
FSM 23	0.40	0.29	0.23	0.23	0.28	0.36	0.18	0.24	0.29	0.40
FSM 24	0.39	0.29	0.27	0.27	0.27	0.37	0.24	0.27	0.28	0.42
FSM 25	0.38	0.31	0.25	0.24	0.33	0.36	0.19	0.23	0.30	0.38
FSM 26	0.47	0.18	0.15	0.18	0.15	0.23	0.10	0.15	0.19	0.41
FSM 27	0.39	0.28	0.26	0.23	0.25	0.29	0.20	0.21	0.27	0.42
FSM 28	0.43	0.27	0.22	0.22	0.20	0.26	0.15	0.17	0.27	0.44
FSM 29	0.37	0.29	0.27	0.25	0.26	0.40	0.23	0.28	0.30	0.41
FSM 30	0.43	0.19	0.15	0.16	0.15	0.21	0.09	0.11	0.20	0.38
FSM 31	0.41	0.22	0.17	0.22	0.22	0.29	0.17	0.17	0.21	0.45
FSM 32	0.42	0.24	0.22	0.23	0.20	0.30	0.17	0.19	0.27	0.43
FSM 33	0.41	0.26	0.22	0.24	0.23	0.31	0.16	0.24	0.25	0.40
FSM 34	0.49	0.24	0.20	0.21	0.20	0.25	0.12	0.16	0.24	0.39
FSM 35	0.47	0.21	0.18	0.26	0.20	0.27	0.11	0.18	0.19	0.43
FSM 36	0.32	0.40	0.35	0.23	0.25	0.39	0.21	0.32	0.31	0.38
FSM 37	0.43	0.29	0.25	0.24	0.24	0.35	0.17	0.22	0.30	0.42
FSM 38	0.36	0.27	0.23	0.24	0.32	0.41	0.20	0.25	0.23	0.38
FSM 39	0.37	0.27	0.24	0.24	0.27	0.34	0.18	0.27	0.24	0.37
FSM 40	0.40	0.29	0.25	0.23	0.28	0.36	0.19	0.26	0.34	0.40
FSM 41	1.00	0.13	0.09	0.11	0.12	0.15	0.06	0.09	0.12	0.35
FSM 42	0.42	1.00	0.44	0.28	0.27	0.37	0.20	0.29	0.35	0.42
FSM 43	0.38	0.49	1.00	0.29	0.28	0.39	0.27	0.32	0.41	0.41
FSM 44	0.37	0.30	0.27	1.00	0.29	0.35	0.25	0.28	0.30	0.44
FSM 45	0.38	0.27	0.25	0.25	1.00	0.39	0.23	0.25	0.25	0.36
FSM 46	0.37	0.29	0.26	0.24	0.30	1.00	0.24	0.26	0.28	0.41
FSM 47	0.31	0.28	0.28	0.29	0.31	0.40	1.00	0.28	0.28	0.37
FSM 48	0.33	0.33	0.31	0.30	0.27	0.37	0.24	1.00	0.33	0.35
FSM 49	0.35	0.33	0.33	0.24	0.21	0.34	0.18	0.27	1.00	0.38
FSM 50	0.44	0.18	0.15	0.21	0.14	0.22	0.09	0.11	0.17	1.00

Appendix J.5 Ratio Model Matrix for Data Set 1, 5 of 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																									■									
2																																												■						
3																																									■									
4																																					■													
5								■																																										
6																■																																		
7	■																																																	
8							■																																											
9							■																																											
10																																									■									
11																																																		■
12																														■																				
13																																	■																	
14																														■																				
15																																														■				
16							■																																											
17																																									■									
18																														■																				
19							■																																											
20											■																																							
21							■																																											
22																■																																		
23																																								■										
24														■																																				
25							■																																											
26																																									■									
27																	■																																	
28																														■																				
29																														■																				
30																																									■									
31							■																																											
32																■																																		
33	■																																																	
34																	■																																	
35																																									■									
36														■																																				
37																														■																				
38												■																																						
39							■																																											
40																							■																											
41	■																																																	
42																																											■							
43																																										■								
44											■																																							
45							■																																											
46							■																																											
47																■																																		
48								■																																										
49							■																																											
50	■																																																	

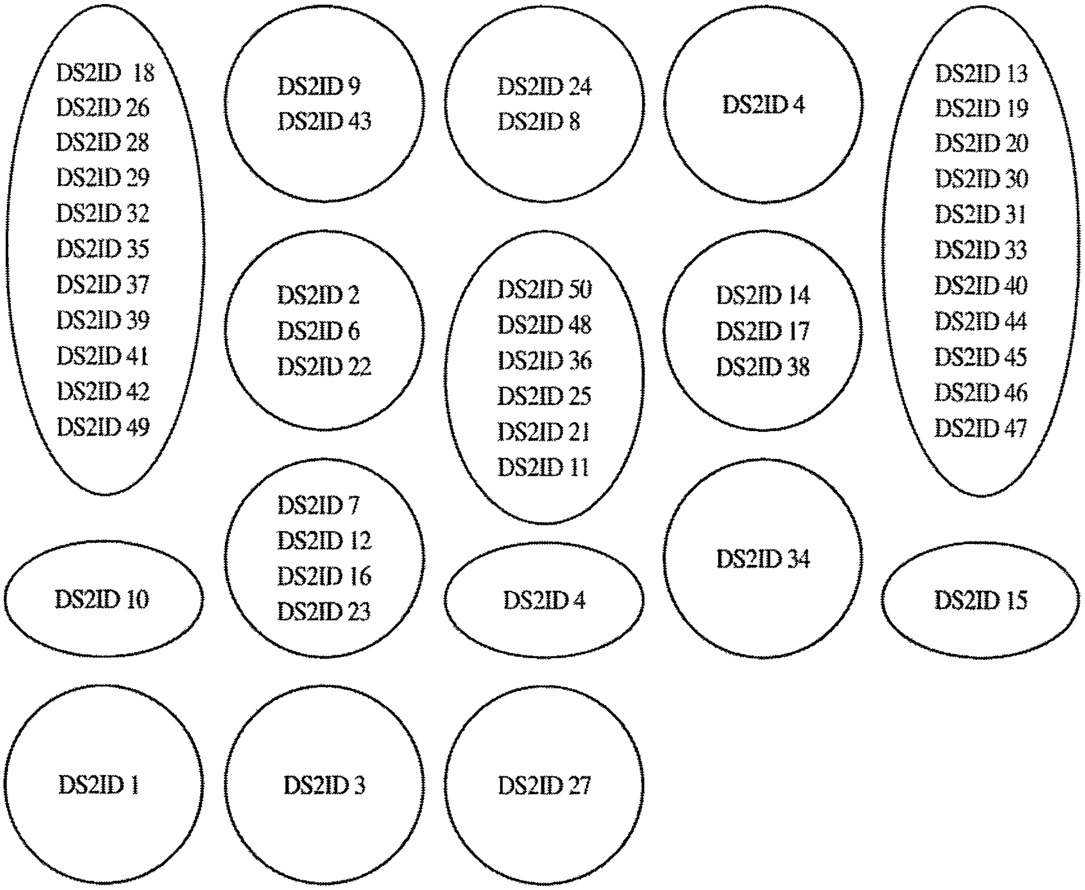
Appendix J.6 Ratio Model Similarity Grid, Data Set 1.



Appendix J.7 Ratio Model Similar Document Clusters, Data Set 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																■																		
2																			■																															
3																																	■																	
4																						■																												
5																		■																																
6																			■																															
7																								■																										
8							■																																											
9																																								■										
10																																																■		
11																															■																			
12																								■																										
13																																											■							
14																														■																				
15														■																																				
16																								■																										
17																														■																				
18	■																																																	
19																																											■							
20																																											■							
21																															■																			
22																			■																															
23																								■																										
24							■																																											
25																															■																			
26	■																																																	
27																																		■																
28	■																																																	
29	■																																																	
30																																											■							
31																																											■							
32	■																																																	
33																																											■							
34																											■																							
35	■																																																	
36																															■																			
37	■																																																	
38																														■																				
39	■																																																	
40																																											■							
41	■																																																	
42	■																																																	
43																																								■										
44																																											■							
45																																											■							
46																																											■							
47																																											■							
48																															■																			
49	■																																																	
50																															■																			

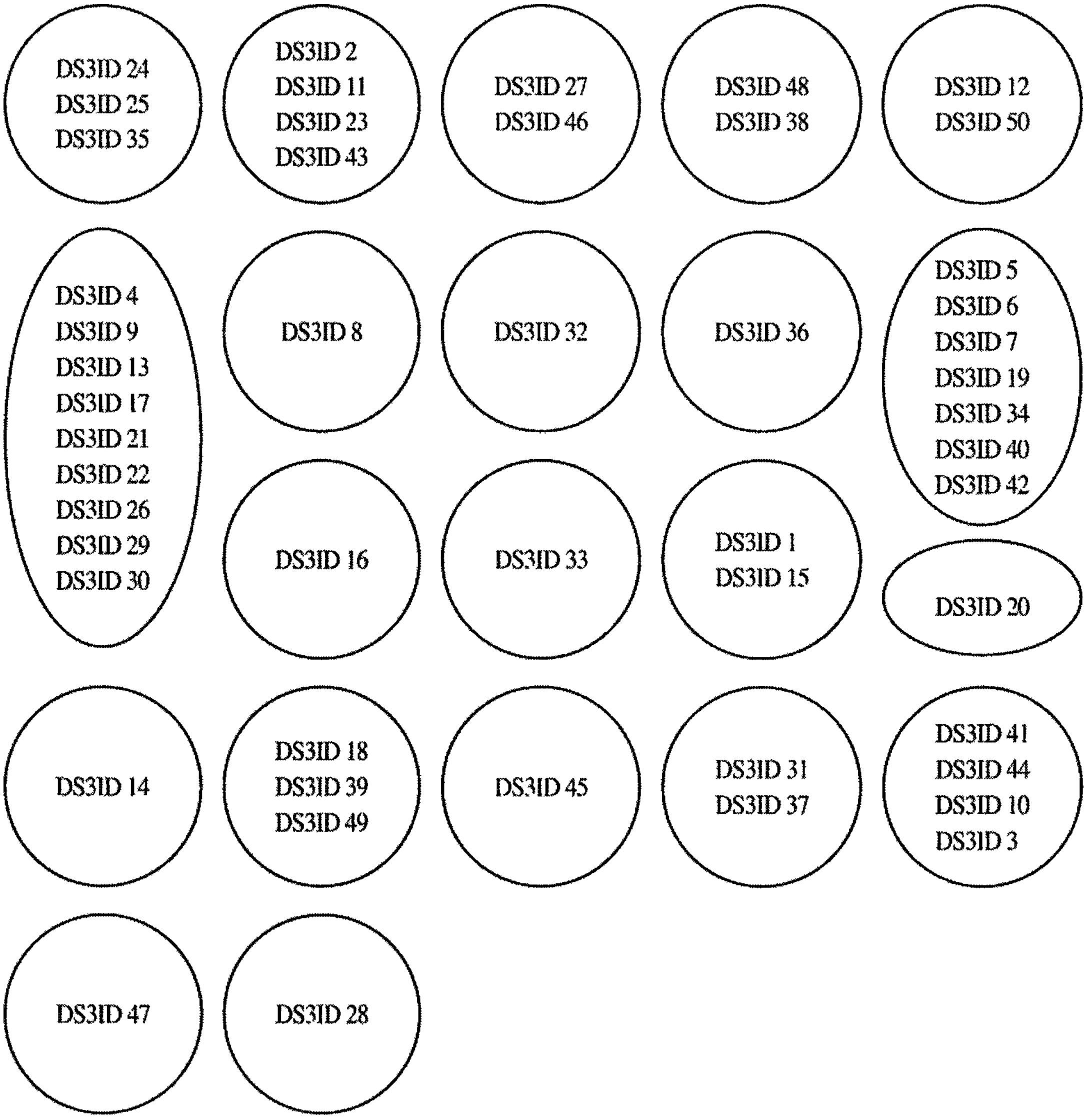
Appendix J.8 Ratio Model Similarity Grid, Data Set 2.



Appendix J.9 Ratio Model Similar Document Clusters, Data Set 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																													■																					
2				■																																														
3																																															■			
4																																																		■
5										■																																								
6										■																																								
7										■																																								
8																					■																													
9																																																		■
10																																															■			
11				■																																														
12																	■																																	
13																																																		■
14																														■																				
15																													■																					
16																									■																									
17																																																		■
18																															■																			
19										■																																								
20															■																																			
21																																																		■
22																																																		■
23				■																																														
24		■																																																
25		■																																																
26																																																		■
27					■																																													
28																																											■							
29																																																		■
30																																																		■
31																																							■											
32																							■																											
33																										■																								
34										■																																								
35		■																																																
36																			■																															
37																																							■											
38						■																																												
39																															■																			
40										■																																								
41																																															■			
42										■																																								
43				■																																														
44																																															■			
45																																■																		
46					■																																													
47																																												■						
48						■																																													
49																															■																			
50																	■																																	

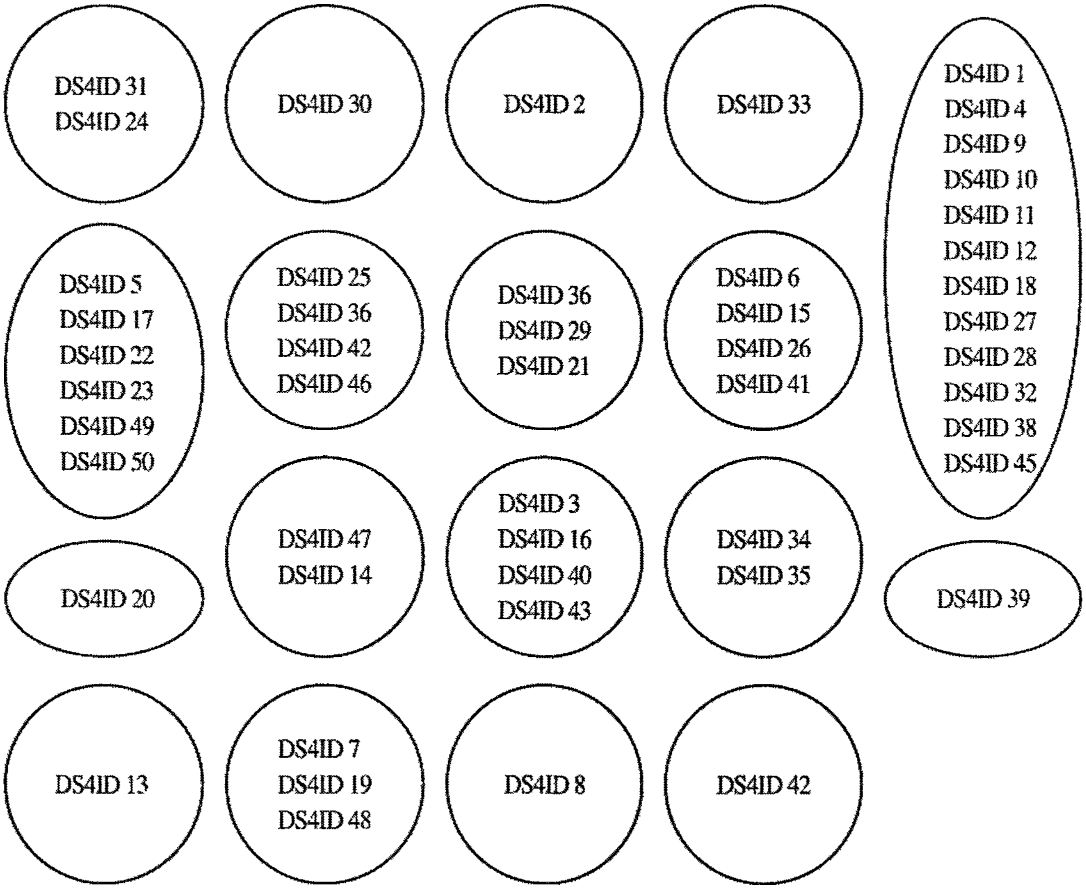
Appendix J.10 Ratio Model Similarity Grid, Data Set 3.



Appendix J.11 Ratio Model Similar Document Clusters, Data Set 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																						■																												
2																																																		■
3													■																																					
4																						■																												
5	■																																																	
6												■																																						
7								■																																										
8																			■																															
9																						■																												
10																						■																												
11																						■																												
12																						■																												
13			■																																															
14											■																																							
15												■																																						
16													■																																					
17	■																																																	
18																						■																												
19								■																																										
20																																				■														
21															■																																			
22	■																																																	
23	■																																																	
24																							■																											
25																																																	■	
26												■																																						
27																						■																												
28																						■																												
29															■																																			
30																											■																							
31																							■																											
32																						■																												
33																																							■											
34										■																																								
35										■																																								
36															■																																			
37																																																	■	
38																						■																												
39																																	■																	
40													■																																					
41												■																																						
42																		■																																
43													■																																				■	
44																																																		
45																						■																												
46																																																	■	
47											■																																							
48								■																																										
49	■																																																	
50	■																																																	

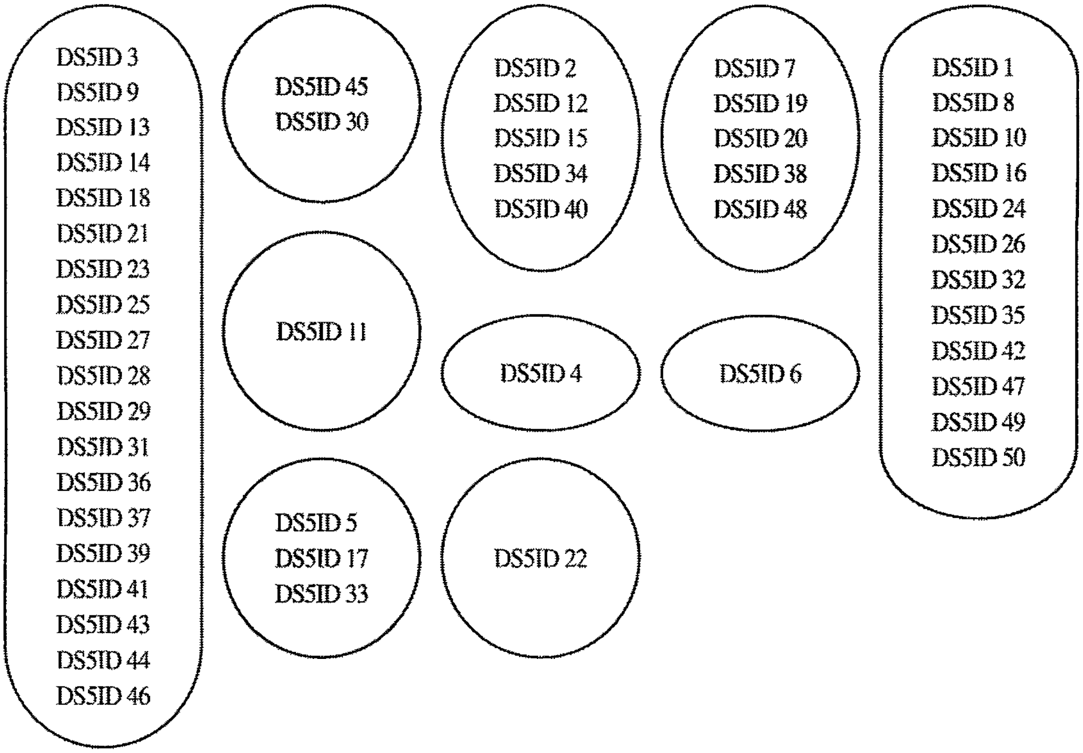
Appendix J.12 Ratio Model Similarity Grid, Data Set 4.



Appendix J.13 Ratio Model Similar Document Clusters, Data Set 4.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																				■														
2																																											■							
3																																		■																
4																																													■					
5																				■																														
6											■																																							
7																					■																													
8																																				■														
9																																		■																
10																																				■														
11						■																																												
12																																											■							
13																																		■																
14																																		■																
15																																											■							
16																																				■														
17																				■																														
18																																		■																
19																					■																													
20																					■																													
21																																		■																
22																							■																											
23																																		■																
24																																				■														
25																																		■																
26																																				■														
27																																		■																
28																																		■																
29																																		■																
30																																								■										
31																																		■																
32																																				■														
33																				■																														
34																																											■							
35	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	■	37	38	39	40	41	42	43	44	45	46	47	48	49	50
36																																		■																
37																																		■																
38																					■																													
39																																		■																
40																																											■							
41																																		■																
42																																				■														
43																																		■																
44																																		■																
45																																								■										
46																																		■																
47																																				■														
48																					■																													
49																																				■														
50																																				■														

Appendix J.15 Ratio Model Similarity Grid, Data Set 5.



Appendix J.16 Ratio Model Similar Document Clusters, Data Set 5.

APPENDIX K

PRODUCT MOMENT EXTENDED RESULTS

DATA SETS 1-5

	List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
FSM 1	1.00	0.05	0.15	-0.31	0.18	0.16	0.01	-0.10	0.06	-0.08
FSM 2	0.13	1.00	0.15	-0.47	-0.02	-0.16	-0.05	-0.15	-0.16	0 09
FSM 3	0.13	-0.13	1.00	-0.23	0.18	0.01	0.10	-0.02	0.01	0.05
FSM 4	0.10	-0.21	0.27	1.00	-0.04	-0.12	0.02	-0.14	-0.17	0.11
FSM 5	0.15	-0.23	0.22	-0.49	1.00	-0.40	-0.20	-0 34	-0.20	-0.10
FSM 6	0.07	-0.27	0.09	-0.40	-0.35	1.00	-0.26	-0.45	-0.07	-0.06
FSM 7	-0.05	-0.36	0.13	-0.49	-0.27	-0.39	1.00	-0.43	-0.22	-0.26
FSM 8	-0.10	-0.33	0.06	-0.54	-0.34	-0.41	-0.36	1.00	-0.15	-0.22
FSM 9	0.08	-0.19	0.13	-0.55	-0.14	-0.08	-0.09	-0.08	1.00	0.01
FSM 10	-0.01	-0.15	0 09	-0.34	-0.30	-0.27	-0.20	-0.29	-0.10	1.00
FSM 11	0.09	-0.25	0.07	-0.38	0.05	-0.04	-0.01	-0.10	-0.03	0.12
FSM 12	-0.08	-0.17	0.03	-0.50	-0.03	-0.13	-0.16	-0.24	-0.08	-0.02
FSM 13	0 08	0.02	0 27	-0.14	0.20	0.21	-0.05	-0.09	-0.03	0.23
FSM 14	-0.02	-0.15	0.06	-0.47	0.07	-0.04	-0.09	-0.03	-0.07	-0.15
FSM 15	0.08	-0.45	-0.14	-0.26	0.02	0.07	-0.01	-0.01	-0.15	-0.04
FSM 16	0.06	-0.17	0.12	-0.46	-0.29	-0.53	-0.23	-0.40	-0.05	-0 18
FSM 17	0.18	-0.30	0.13	-0.36	0.07	0.08	0.14	0.03	0.12	-0.09
FSM 18	0.00	-0.05	0.19	-0.24	0.12	-0.21	-0.01	-0.23	0.01	0.11
FSM 19	0.02	-0.23	0.12	-0.46	-0.04	-0.08	-0.13	-0.12	-0.02	0.09
FSM 20	0.11	-0.19	0.22	-0.34	-0.33	-0.44	-0.25	-0.47	-0.20	-0.06
FSM 21	0.05	-0.31	0.09	-0.34	-0.12	-0.14	-0.08	-0.21	-0.19	0.04
FSM 22	0.07	-0.02	0 19	-0.40	-0.13	0.09	0.01	0.02	0.18	0.00
FSM 23	0.02	-0.05	0.28	-0.19	0.18	0.03	0.08	0.14	0.15	0.21
FSM 24	0.09	-0.15	0.15	-0.22	0.21	0.14	0.09	-0.01	-0.12	0.16
FSM 25	0.10	-0.01	0.20	-0.19	0.12	0.18	0.03	0.04	0.09	0.13
FSM 26	-0.09	-0.25	0 22	-0.54	-0.28	-0.45	-0.25	-0.36	-0.17	-0.18
FSM 27	0.15	0.03	0.29	-0 07	0.17	0.15	0.17	0.14	0.12	0.23
FSM 28	-0.00	0.05	0.07	-0.17	-0.02	-0.00	-0.01	-0.07	-0.05	-0.02
FSM 29	-0.08	-0.09	0.05	-0.31	0.15	0.06	0.04	0.04	0.11	0.10
FSM 30	-0.15	-0.02	-0.09	-0.11	0.12	-0.01	0.01	-0.17	-0.00	-0.06
FSM 31	0 02	-0.23	0.12	-0.46	-0.04	-0.08	-0.13	-0.12	-0.02	0.09
FSM 32	-0.03	-0.16	0.10	-0.45	-0.07	-0.24	-0.09	-0.26	-0.01	-0.01
FSM 33	0.05	-0.00	0 22	-0.13	0.18	0.18	0.03	-0.08	0.10	0.15
FSM 34	0.10	-0.22	0.09	-0.41	-0.03	0.04	0.04	-0.10	0.04	-0.06
FSM 35	-0.05	-0.17	0.09	-0.33	-0.22	-0.26	-0.21	-0.30	0.03	-0.16
FSM 36	-0.04	-0.23	-0.11	-0.45	0.09	0.16	-0.07	-0.11	-0.28	-0.03
FSM 37	-0.03	-0.21	-0.01	-0.28	-0.03	-0.11	0 02	-0.08	-0.03	-0.03
FSM 38	0.10	0.02	0.08	-0.21	0.13	0.01	0.09	0.05	0.06	0.19
FSM 39	0 05	-0.10	0.18	-0.42	0.05	-0.08	-0.03	-0.07	0.02	0.07
FSM 40	0.14	-0.11	0.18	-0.16	0.22	0.10	0.04	0.07	0.06	0.17
FSM 41	0 16	-0.34	-0.01	-0.46	-0.25	-0.31	-0.04	-0.29	-0.16	-0.11
FSM 42	0.08	-0 10	0.10	-0.26	0.10	0.14	0.07	-0.06	-0.11	0.02
FSM 43	0 07	0.02	0.26	-0.35	0.13	0.13	0.13	-0.01	0.10	0.09
FSM 44	0.10	-0.30	0 12	-0.48	-0.12	-0.24	-0.11	-0.20	-0.07	0.05
FSM 45	0.23	-0 26	0.28	-0.32	0.05	-0.31	0.12	0.01	0.00	0.25
FSM 46	0.13	-0.32	0.01	-0.36	-0.02	-0.03	-0.03	-0.06	-0.15	0.11
FSM 47	0.14	-0.28	0.26	-0.36	-0.02	-0.15	0.01	-0.08	-0.11	0.18
FSM 48	0.08	-0.31	0.15	-0.33	-0.30	-0.47	-0.27	-0.45	-0.13	-0.14
FSM 49	0.02	-0.20	0.00	-0.14	0.10	-0.14	-0.05	-0.14	-0.11	-0.03
FSM 50	-0.08	-0.17	0.14	-0.36	0.20	0.11	-0.10	0.00	-0.01	0 04

Appendix K 1 Product Moment Matrix for Data Set 1, 1 of 5

	List 11	List 12	List 13	List 14	List 15	List 16	List 17	List 18	List 19	List 20
FSM 1	-0.01	-0.11	-0.29	-0.07	0.19	0.10	0.11	-0.13	-0.07	0.16
FSM 2	-0 31	-0.06	0.14	-0.26	-0.22	-0.03	-0.09	0.15	-0.14	-0.16
FSM 3	-0.23	-0.03	-0.05	0.00	-0.14	-0.01	0.03	0.03	0.01	0.17
FSM 4	-0.08	-0.05	-0.04	-0.16	0 20	-0.03	0 05	0.06	-0.07	0.04
FSM 5	-0.25	-0.08	-0.09	-0.02	-0.00	-0.28	-0.02	0.00	-0.13	-0.51
FSM 6	-0.27	-0.10	0.08	-0.11	0.06	-0.51	0.01	-0.16	-0.10	-0.59
FSM 7	-0.34	-0.23	-0.47	-0.20	0.02	-0.28	0.00	-0.20	-0.31	-0.56
FSM 8	-0.39	-0.27	-0.32	-0.16	0.19	-0.39	-0.16	-0.36	-0.25	-0.56
FSM 9	-0.10	0.01	-0.23	-0.13	0.10	0.00	0.02	-0.01	-0.07	-0.22
FSM 10	0.01	-0.05	-0.10	-0.24	-0.12	-0.23	-0.15	-0.14	-0.00	-0.25
FSM 11	1.00	-0.23	0.09	-0.06	0.05	0.06	0.06	0.20	-0.10	0.10
FSM 12	-0.43	1.00	-0.33	0 04	0.17	-0.15	-0.01	-0.16	-0.20	-0.14
FSM 13	-0.01	-0.03	1.00	0.04	0.17	0.14	0.12	0.02	0.09	0.22
FSM 14	-0.24	-0.06	-0.05	1.00	-0.04	-0.11	-0.27	-0.22	-0.13	0.06
FSM 15	-0.35	0.01	0.00	-0.23	1.00	-0.04	-0.17	-0.09	-0.12	-0.07
FSM 16	-0.14	-0.18	-0.07	-0.08	0.14	1.00	-0.02	-0.22	-0.06	-0.57
FSM 17	-0.03	0.05	0.01	-0.24	-0.18	0.00	1.00	-0.22	0.01	0.15
FSM 18	-0.05	-0.01	-0.14	-0.13	0.11	-0.13	-0.08	1.00	0.06	0.08
FSM 19	-0.21	-0.13	-0.05	-0.11	0.10	-0.01	0.01	0.11	1.00	-0.13
FSM 20	-0.15	-0.11	0.02	-0.04	0.07	-0.32	0.02	0.07	-0.09	1.00
FSM 21	-0.31	-0.06	0.00	-0.32	0.09	-0.08	0.01	0.07	-0.06	-0.14
FSM 22	-0.08	0.06	0.00	-0.23	0.17	-0.01	-0.05	0.06	0.02	0.28
FSM 23	-0.18	0.09	0.13	-0.02	0.14	0.03	0.08	0 05	-0.07	0.26
FSM 24	-0.20	0.19	0.20	-0 08	0 25	0.03	0.06	0.20	-0.13	0.23
FSM 25	-0.18	0.03	0.17	-0.26	0.15	0.08	-0.05	0.14	-0.00	0.17
FSM 26	-0.25	-0.17	-0.10	-0.22	0.02	-0.30	-0.03	-0.07	-0.17	-0.54
FSM 27	-0.09	0.07	0.10	0 02	0.12	0.11	0.09	0.07	0.08	0.22
FSM 28	-0.01	0.01	-0.04	-0.24	-0.13	-0.08	-0.16	-0.07	-0.01	0.10
FSM 29	-0.29	0.02	0.11	-0.29	0.07	-0.00	-0.01	-0 17	-0.07	0.15
FSM 30	-0.16	-0.14	0.02	-0.17	0.10	-0.13	-0.18	-0.47	-0.12	0.13
FSM 31	-0.21	-0.13	-0.05	-0.11	0.10	-0.01	0.01	0.11	1.00	-0.13
FSM 32	-0.09	-0.28	-0.12	-0.06	0.06	-0.14	-0.04	-0.14	-0.15	-0.27
FSM 33	-0.40	-0.26	-0.16	-0.06	0.30	0.08	0 15	-0.09	0.06	0.23
FSM 34	-0.21	0.06	-0.04	-0.30	-0.03	-0.06	-0.16	-0.07	-0.03	0.06
FSM 35	-0 19	-0.17	-0.21	-0.22	0.14	-0.16	-0.02	-0.15	-0.13	-0.33
FSM 36	-0.42	-0.14	-0 19	-0.23	0.10	-0.15	-0.14	-0.08	-0.17	0.02
FSM 37	-0.25	-0.12	-0.09	-0.29	0.04	-0.12	-0.13	-0.09	-0.17	0.09
FSM 38	-0.28	-0.03	0.21	-0.08	-0.11	0.03	0.05	0 17	0.04	0.24
FSM 39	-0.23	-0.10	-0.05	-0.05	0 11	-0.08	0.01	-0.25	-0.05	-0.01
FSM 40	-0.22	0.05	0.01	-0.03	0.17	0.06	-0.01	0.10	0.02	0.03
FSM 41	-0 20	-0.15	-0.10	-0.14	-0.09	-0.22	0.07	-0.12	-0.15	-0.59
FSM 42	-0.06	0.08	-0.02	-0.21	-0.13	0.01	-0.05	0.01	-0.06	0.14
FSM 43	-0.07	0.05	0.01	-0.27	-0.09	0.07	-0.03	0.13	0.14	0.13
FSM 44	-0.25	-0.05	-0.01	-0.13	0.13	-0.18	-0.04	-0.04	-0.15	-0.09
FSM 45	-0.28	0.02	0.30	0.06	0.21	-0.17	0.14	0.16	-0.05	-0.08
FSM 46	-0.39	-0.08	0.03	-0.11	0.03	-0.07	-0.00	-0.06	-0.17	-0.16
FSM 47	-0.25	-0.05	0.19	-0.07	0.07	-0.07	0.03	0.15	-0.12	-0.15
FSM 48	-0.11	-0.11	-0.14	-0.07	-0.16	-0.34	-0.11	-0.11	-0.03	-0.57
FSM 49	-0.22	-0.09	-0.00	-0.29	-0.18	-0.17	-0.17	-0.12	-0.01	-0.04
FSM 50	-0.15	-0.01	0.10	-0.14	0 05	-0.02	0.07	0.11	-0.23	0.21

Appendix K.2 Product Moment Matrix for Data Set 1, 2 of 5.

	List 21	List 22	List 23	List 24	List 25	List 26	List 27	List 28	List 29	List 30
FSM 1	-0.18	0.00	-0.15	0.08	-0.02	-0.05	0.01	-0.07	-0.24	-0.17
FSM 2	-0.22	-0.05	0.25	0.00	-0.06	0.04	0.15	0.07	0.23	0.08
FSM 3	0.14	0.11	0.17	0.04	0.07	0.14	0.22	-0.05	-0.11	-0.08
FSM 4	-0.18	-0.10	0.16	0.27	-0.01	0.07	0.16	0.03	0.05	0.18
FSM 5	-0.40	-0.25	0.10	0.19	-0.09	-0.09	0.10	-0.10	0.07	0.05
FSM 6	-0.31	-0.06	-0.14	0.07	0.07	-0.24	0.08	0.05	0.01	-0.02
FSM 7	-0.36	-0.13	-0.01	0.00	-0.20	-0.22	0.08	-0.16	-0.03	-0.06
FSM 8	-0.34	-0.07	0.01	0.01	-0.06	-0.24	-0.13	-0.22	-0.01	-0.15
FSM 9	-0.29	0.11	0.17	-0.16	-0.04	0.00	0.09	-0.09	0.12	0.09
FSM 10	-0.19	-0.17	0.17	0.05	-0.12	-0.12	0.06	-0.11	0.04	-0.06
FSM 11	-0.27	-0.02	0.13	0.08	-0.13	0.08	0.01	0.04	-0.05	0.07
FSM 12	-0.21	0.02	0.17	0.34	-0.00	-0.03	-0.05	-0.13	-0.07	-0.09
FSM 13	0.21	0.03	0.11	0.29	0.10	0.23	0.13	0.04	0.12	0.13
FSM 14	-0.22	-0.26	0.06	0.08	-0.32	-0.09	-0.03	-0.25	-0.30	-0.11
FSM 15	0.07	0.02	-0.11	0.04	-0.05	0.02	0.11	-0.19	-0.07	-0.12
FSM 16	-0.24	-0.05	0.06	0.06	0.05	-0.21	-0.01	-0.11	0.07	-0.12
FSM 17	0.13	-0.07	0.22	-0.01	-0.26	0.03	0.02	-0.23	0.03	-0.12
FSM 18	0.14	-0.06	0.00	0.18	0.03	0.09	0.03	-0.05	-0.16	-0.13
FSM 19	-0.23	-0.01	-0.05	-0.16	0.02	0.05	-0.00	0.01	0.02	-0.00
FSM 20	-0.29	0.11	0.07	0.07	0.03	-0.16	0.02	0.08	0.13	0.09
FSM 21	1.00	0.05	0.08	0.04	-0.00	0.02	0.05	0.01	0.06	-0.11
FSM 22	0.14	1.00	0.04	0.23	-0.12	0.04	-0.05	-0.14	0.15	-0.00
FSM 23	0.07	0.06	1.00	0.33	0.07	0.06	-0.20	0.03	0.17	0.10
FSM 24	0.13	0.09	0.24	1.00	0.05	0.16	0.11	0.01	0.14	0.05
FSM 25	0.07	-0.13	0.18	0.19	1.00	0.20	0.05	0.12	0.07	-0.01
FSM 26	-0.23	-0.12	-0.17	0.17	-0.06	1.00	0.03	-0.24	-0.02	-0.02
FSM 27	0.08	0.10	0.02	0.22	0.02	0.24	1.00	0.06	0.14	0.10
FSM 28	0.01	-0.17	0.01	0.03	0.01	-0.03	-0.12	1.00	-0.05	-0.04
FSM 29	0.01	0.09	0.01	0.10	0.01	0.09	0.07	-0.01	1.00	-0.15
FSM 30	-0.07	-0.13	0.14	0.00	0.00	0.03	-0.00	-0.17	-0.43	1.00
FSM 31	-0.23	-0.01	-0.05	-0.16	0.02	0.05	-0.00	0.01	0.02	-0.00
FSM 32	-0.32	-0.09	0.02	0.16	-0.11	-0.05	-0.07	-0.13	-0.04	-0.11
FSM 33	0.28	0.04	0.16	0.33	0.16	0.21	0.05	0.02	0.12	0.04
FSM 34	0.05	-0.15	0.03	0.11	-0.06	0.03	0.02	-0.20	0.03	-0.01
FSM 35	0.14	0.01	0.17	0.01	-0.04	-0.09	0.09	-0.09	-0.22	-0.07
FSM 36	-0.05	-0.10	-0.04	-0.09	-0.03	-0.03	-0.13	-0.25	-0.05	-0.15
FSM 37	0.01	-0.07	0.10	0.08	-0.04	0.02	0.00	-0.11	-0.44	-0.02
FSM 38	0.09	0.03	0.19	0.09	-0.09	0.13	0.13	0.12	-0.19	0.05
FSM 39	0.06	-0.01	-0.10	0.19	0.04	0.07	0.04	-0.01	0.05	0.04
FSM 40	0.08	0.04	-0.25	0.07	0.02	0.11	-0.23	-0.11	0.26	0.07
FSM 41	-0.39	-0.21	0.04	0.10	-0.14	-0.13	-0.11	-0.13	0.05	-0.02
FSM 42	0.17	0.03	0.02	0.14	0.02	0.14	0.08	-0.12	-0.01	-0.02
FSM 43	0.13	0.04	0.07	0.23	0.09	0.13	-0.00	-0.11	0.03	0.10
FSM 44	-0.29	-0.09	0.10	0.17	0.03	0.03	0.10	-0.07	0.01	0.10
FSM 45	-0.20	0.06	0.11	0.24	-0.13	0.16	0.13	0.16	0.22	0.13
FSM 46	-0.39	0.06	0.12	0.17	0.01	-0.02	0.10	0.01	-0.18	-0.07
FSM 47	-0.32	0.05	0.08	0.02	0.02	0.08	-0.07	0.05	0.13	0.10
FSM 48	-0.06	-0.14	-0.06	0.13	0.12	-0.20	0.14	-0.08	-0.05	0.05
FSM 49	-0.07	-0.22	-0.00	0.13	-0.01	-0.07	-0.04	-0.29	-0.02	-0.21
FSM 50	-0.16	0.11	0.11	-0.13	-0.02	0.04	0.04	-0.01	-0.04	0.01

Appendix K 3 Product Moment Matrix for Data Set 1, 3 of 5

	List 31	List 32	List 33	List 34	List 35	List 36	List 37	List 38	List 39	List 40
FSM 1	-0.07	-0.04	-0.12	0.08	-0.21	-0.17	-0.06	0.04	-0.10	0.16
FSM 2	-0.14	-0.05	0.02	0.09	-0.15	-0.15	0.02	0.05	0.03	0.15
FSM 3	0.01	0.02	-0.06	0.09	-0.15	-0.15	-0.13	-0.14	0.04	0.04
FSM 4	-0.07	-0.04	-0.03	0.05	-0.01	-0.26	0.05	0.06	-0.18	0.14
FSM 5	-0.13	-0.15	-0.03	0.06	-0.30	-0.13	-0.04	-0.04	-0.18	0.04
FSM 6	-0.10	-0.24	0.02	0.11	-0.28	0.04	0.01	-0.03	-0.14	-0.04
FSM 7	-0.31	-0.19	-0.15	-0.02	-0.39	-0.23	-0.02	0.01	-0.20	-0.07
FSM 8	-0.25	-0.36	-0.21	-0.03	-0.38	-0.26	-0.05	0.08	-0.27	-0.07
FSM 9	-0.07	0.05	0.00	0.16	-0.06	-0.34	0.01	0.01	-0.03	0.12
FSM 10	-0.00	-0.05	-0.01	-0.10	-0.30	-0.13	-0.04	0.14	-0.18	0.13
FSM 11	-0.10	0.04	-0.28	0.07	-0.16	-0.28	0.01	-0.22	0.03	0.12
FSM 12	-0.20	-0.35	-0.38	0.03	-0.23	-0.31	-0.14	-0.06	-0.35	0.04
FSM 13	0.09	0.14	-0.12	0.19	-0.11	-0.10	0.06	0.14	0.00	0.04
FSM 14	-0.13	-0.12	-0.20	-0.15	-0.25	-0.18	-0.22	-0.07	-0.05	0.07
FSM 15	-0.12	-0.07	0.01	-0.07	-0.18	-0.02	-0.07	-0.34	-0.08	-0.01
FSM 16	-0.06	-0.19	0.01	0.05	-0.29	-0.22	-0.07	-0.06	-0.26	-0.01
FSM 17	0.01	-0.04	0.07	-0.08	-0.12	-0.17	-0.08	0.03	-0.02	0.14
FSM 18	0.06	0.05	-0.13	0.11	-0.12	-0.03	-0.02	0.09	-0.21	0.11
FSM 19	1.00	-0.07	-0.05	0.08	-0.18	-0.17	-0.10	-0.13	-0.08	0.11
FSM 20	-0.09	-0.17	0.05	0.13	-0.25	-0.08	0.10	0.10	-0.07	-0.04
FSM 21	-0.06	-0.17	0.09	0.11	-0.08	-0.15	0.01	0.01	0.11	0.01
FSM 22	0.02	-0.01	0.04	0.02	-0.05	-0.13	0.02	0.13	-0.00	0.08
FSM 23	-0.07	0.04	-0.00	0.13	-0.06	0.00	0.17	0.27	-0.06	-0.21
FSM 24	-0.13	0.10	0.15	0.13	-0.10	-0.07	0.12	0.16	0.12	0.15
FSM 25	-0.00	-0.03	0.19	0.15	-0.07	0.03	0.14	-0.09	0.14	0.12
FSM 26	-0.17	-0.19	-0.07	-0.03	-0.28	-0.27	-0.05	-0.06	-0.21	-0.09
FSM 27	0.08	0.04	0.12	0.17	0.02	0.08	0.21	0.12	0.16	-0.08
FSM 28	-0.01	-0.06	-0.08	-0.13	-0.15	-0.29	-0.02	-0.02	-0.06	-0.13
FSM 29	-0.07	-0.00	0.02	0.12	-0.23	-0.14	-0.29	-0.33	-0.04	0.19
FSM 30	-0.12	-0.20	-0.17	-0.00	-0.19	-0.19	-0.07	-0.00	-0.05	0.07
FSM 31	1.00	-0.07	-0.05	0.08	-0.18	-0.17	-0.10	-0.13	-0.08	0.11
FSM 32	-0.15	1.00	-0.06	-0.09	-0.21	-0.22	-0.11	0.02	-0.24	0.01
FSM 33	0.06	0.13	1.00	0.17	-0.12	-0.26	0.00	0.12	-0.14	0.16
FSM 34	-0.03	-0.13	-0.00	1.00	-0.14	-0.39	0.01	0.21	-0.13	0.03
FSM 35	-0.13	-0.14	-0.06	0.11	1.00	-0.30	-0.23	-0.03	-0.11	0.06
FSM 36	-0.17	-0.15	-0.21	-0.19	-0.23	1.00	-0.14	-0.15	-0.34	-0.11
FSM 37	-0.17	-0.09	-0.22	0.02	-0.31	-0.26	1.00	-0.33	0.04	0.05
FSM 38	0.04	0.13	0.12	0.18	-0.05	-0.13	-0.11	1.00	0.13	0.12
FSM 39	-0.05	-0.06	-0.12	0.04	-0.13	-0.35	0.08	0.06	1.00	-0.07
FSM 40	0.02	0.08	0.02	0.07	-0.11	-0.12	0.12	0.10	-0.06	1.00
FSM 41	-0.15	-0.15	0.07	0.08	-0.30	-0.28	0.05	-0.25	-0.13	0.10
FSM 42	-0.06	-0.01	0.02	-0.04	-0.13	-0.30	-0.05	0.18	-0.03	0.10
FSM 43	0.14	-0.05	0.03	-0.01	-0.13	-0.20	-0.13	0.10	-0.01	0.15
FSM 44	-0.15	-0.20	0.00	0.06	-0.19	-0.16	-0.02	0.13	-0.09	0.13
FSM 45	-0.05	0.04	0.18	0.27	-0.04	0.06	0.19	-0.01	0.11	0.11
FSM 46	-0.17	-0.10	0.06	0.01	-0.16	-0.13	-0.10	-0.35	0.10	0.07
FSM 47	-0.12	-0.09	0.05	0.18	-0.02	0.11	0.18	-0.01	0.18	0.17
FSM 48	-0.03	-0.12	-0.14	0.07	-0.31	-0.25	-0.08	0.06	-0.21	-0.05
FSM 49	-0.01	-0.13	-0.04	-0.05	-0.20	-0.08	-0.11	0.01	0.11	0.06
FSM 50	-0.23	0.03	-0.03	0.11	-0.23	-0.25	-0.11	-0.23	0.05	0.06

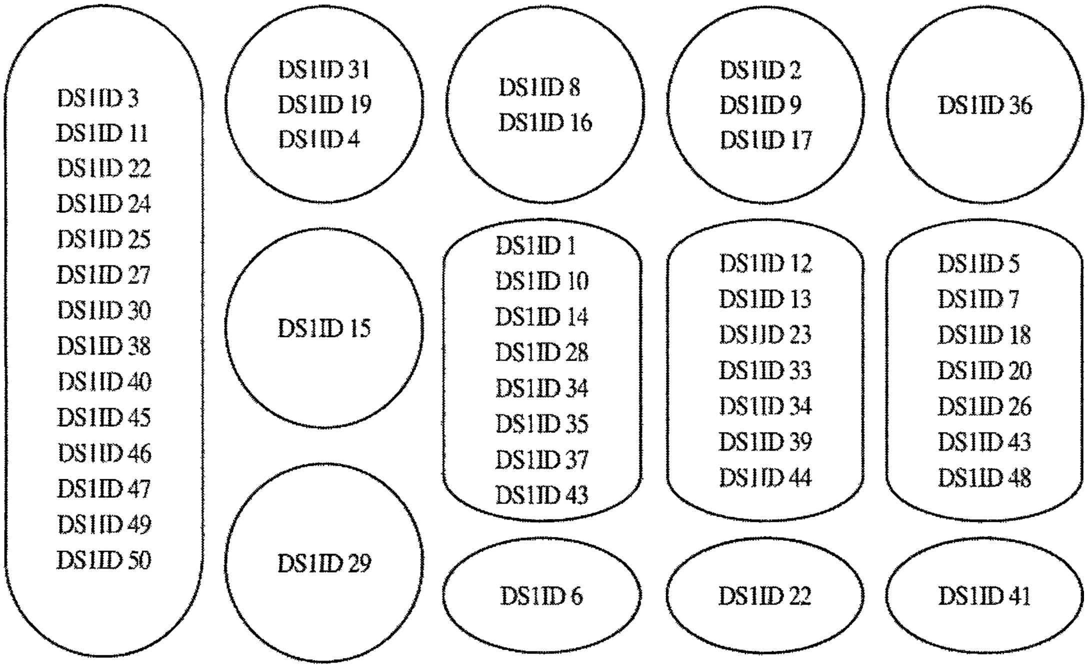
Appendix K 4 Product Moment Matrix for Data Set 1, 4 of 5.

	List 41	List 42	List 43	List 44	List 45	List 46	List 47	List 48	List 49	List 50
FSM 1	0.12	-0.10	-0.14	0.14	0.23	0.13	-0.13	0.15	0.03	0.04
FSM 2	0.08	0.10	0.23	-0.26	-0.19	-0.23	-0.39	-0.05	0.10	-0.01
FSM 3	0.10	-0.05	0.10	0.01	0.11	-0.12	-0.10	0.24	-0.06	0.09
FSM 4	0.18	0.03	-0.12	-0.16	-0.07	-0 07	-0 35	0.13	0.35	0 10
FSM 5	0.04	-0.03	0 01	-0.31	-0.16	-0.09	-0.37	-0.22	0.11	0 06
FSM 6	-0.01	0.01	0.12	-0.35	-0.46	-0.17	-0.34	-0.37	-0.07	-0.01
FSM 7	-0.01	-0.19	-0.12	-0.37	-0.11	-0.08	-0.46	-0.33	-0.00	-0.11
FSM 8	-0.07	-0.16	-0.19	-0.41	-0.14	-0.11	-0.42	-0.38	-0.04	-0.05
FSM 9	0.05	-0.14	-0.01	-0.08	-0.07	-0.10	-0.44	0.00	0.02	0.05
FSM 10	0.04	-0.20	-0.26	-0.06	0.30	0.04	-0.17	-0.27	-0.13	0.13
FSM 11	0.15	0.12	0.10	-0.14	-0.15	-0.26	-0.40	0.41	0.23	0 04
FSM 12	0.06	-0 02	-0.15	-0.12	-0.12	-0.13	-0.47	-0.01	-0.05	0.04
FSM 13	0.22	0.00	-0.00	-0.03	0.19	0.15	-0.05	0.23	0.17	0.17
FSM 14	0.02	-0.25	-0.35	-0.06	0.22	0.01	-0.16	0.20	-0.32	-0.03
FSM 15	0.00	-0.26	-0.24	-0.15	-0.05	-0.10	-0.21	0.01	-0.24	-0.07
FSM 16	-0.00	-0.04	-0.05	-0.30	-0.47	-0.05	-0.39	-0.35	-0.04	0.04
FSM 17	0.15	-0.21	-0.20	-0.03	0.15	0.13	-0.25	-0.05	-0.15	0.11
FSM 18	0.14	0.04	0.04	-0.10	0.12	0.00	-0.04	0.12	-0.02	0.05
FSM 19	0.08	-0.12	0 04	-0.23	-0.14	-0.21	-0 42	0.17	0.25	-0.07
FSM 20	-0.03	0.08	0.11	-0.15	-0.15	-0.16	-0.37	-0.39	0.07	0.11
FSM 21	0.09	-0.03	0 04	-0.29	-0.15	-0.27	-0 42	0.20	0.00	-0.04
FSM 22	0.11	0.07	0.04	0.03	0.13	0.09	-0.15	0.15	0.02	0.16
FSM 23	0.22	0.02	0 07	-0.05	0.20	0 04	-0.02	0.21	0.10	0.01
FSM 24	0.21	0.06	0 12	0 05	0.13	0.08	-0.20	0.46	0.21	-0.04
FSM 25	0.16	0.09	0.07	0.06	-0.09	0.11	-0.11	0.49	0.11	0.13
FSM 26	0.00	-0.03	-0.18	-0.17	-0.07	-0.15	-0.39	-0.37	-0.24	0 04
FSM 27	0.27	0.16	0 12	0.07	0.21	-0.02	-0 14	0.45	0 13	0.14
FSM 28	0.08	-0.21	-0.28	-0.08	0.17	0.03	-0.17	0.08	-0.38	0.10
FSM 29	0.11	-0.07	-0.06	-0.18	0.12	-0.22	-0.05	0.18	0.16	-0.05
FSM 30	0 08	-0.05	-0.05	0 05	0.14	-0.15	-0 19	0.19	-0.33	0.08
FSM 31	0.08	-0.12	0.04	-0.23	-0.14	-0.21	-0.42	0.17	0.25	-0.07
FSM 32	0.06	-0.13	-0.21	-0.27	-0.12	-0.10	-0.45	-0.09	-0.12	0.13
FSM 33	0.22	0.01	0.05	-0.01	0.20	0.23	-0.15	0.18	0.23	0.13
FSM 34	0.09	-0.23	-0.24	-0.20	0.27	-0.11	-0.09	0.07	-0.14	0.04
FSM 35	0.02	-0.11	-0.18	-0.09	0.20	0.05	-0.15	-0.27	-0.09	-0.01
FSM 36	-0.04	-0.29	-0.21	-0.24	0.06	-0.13	-0.20	0.10	-0.04	-0 13
FSM 37	0.03	-0.21	-0.33	-0.25	0.10	-0.21	-0.15	0.04	-0.18	-0.07
FSM 38	0.13	0.08	0 13	0.12	0.00	-0.28	-0.04	0.30	0.25	-0.04
FSM 39	0.13	-0.06	-0.04	-0.18	0.14	0.06	-0.10	-0.03	0.12	0.00
FSM 40	0.18	0.05	0.13	-0.07	0.15	0.10	-0.10	0.34	0.10	0.04
FSM 41	1 00	-0.29	-0.10	-0.24	0.04	-0.00	-0.40	-0.38	-0.18	-0.01
FSM 42	0.07	1.00	-0.39	-0.12	0.23	0.05	-0.06	0 08	-0.14	0.07
FSM 43	0 19	-0.30	1.00	-0.15	0.18	0.08	-0.06	0.16	-0.14	0.09
FSM 44	0.12	-0.01	-0.05	1 00	-0.10	-0.16	-0.39	0.00	-0.05	-0.02
FSM 45	0.19	0.13	0.17	-0.22	1.00	-0.09	-0 37	0.51	0.25	0.05
FSM 46	0.08	-0.10	0.02	-0.27	-0.13	1.00	-0.38	0.30	0 00	-0.07
FSM 47	0.15	0.14	0.16	-0.23	-0.15	-0.16	1.00	0.28	0.18	0.04
FSM 48	-0.07	-0.17	-0.09	-0.23	0.10	-0.07	-0.10	1.00	-0.05	0.03
FSM 49	0.04	-0.23	-0.31	-0.27	0 12	0.01	-0.13	0.21	1.00	-0.00
FSM 50	0.09	0.03	-0.06	-0.04	0.16	-0 09	-0 25	0.26	0.11	1.00

Appendix K.5 Product Moment Matrix for Data Set 1, 5 of 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																													■					
2																							■																											
3																																																■		
4																																																	■	
5			■																																															
6																																											■							
7			■																																															
8															■																																			
9																							■																											
10																																													■					
11																																																■		
12																								■																										
13																								■																										
14																																													■					
15																											■																							
16															■																																			
17																							■																											
18			■																																															
19																																																	■	
20			■																																															
21																																																■		
22																				■																														
23																								■																										
24																																																■		
25																																																■		
26			■																																															
27																																																■		
28																																													■					
29																																								■										
30																																																■		
31																																																	■	
32																								■																										
33																								■																										
34																																													■					
35																																													■					
36						■																																												
37																																													■					
38																																																■		
39																								■																										
40																																																■		
41	■																																																	
42																																													■					
43			■																																															
44																								■																										
45																																																■		
46																																																■		
47																																																■		
48			■																																															
49																																																■		
50																																																■		

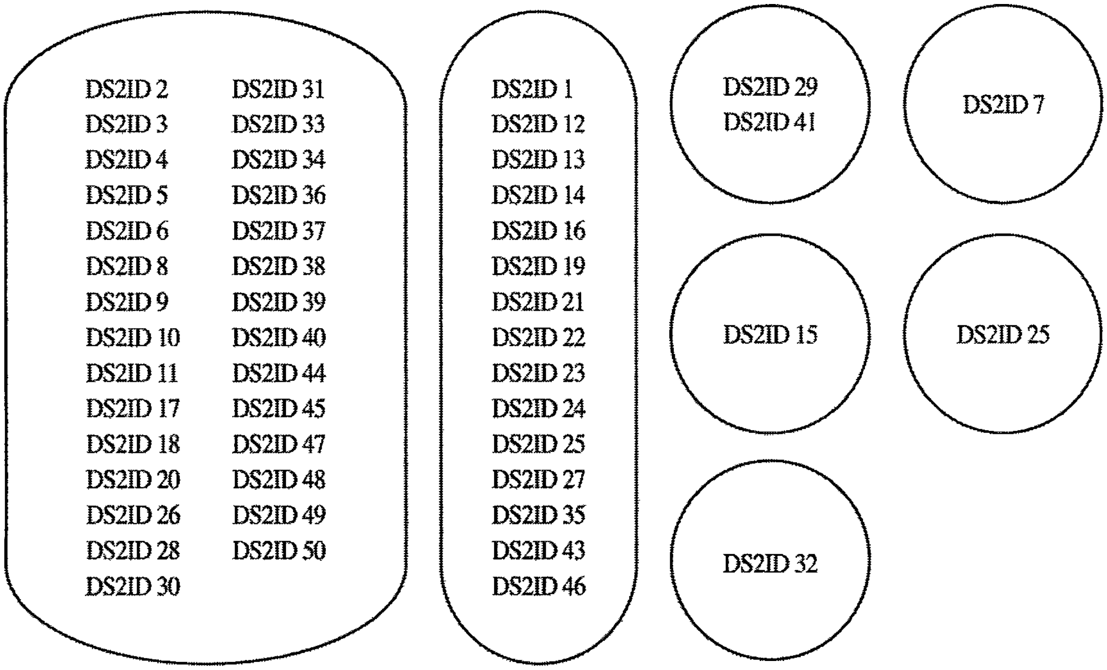
Appendix K.6 Product Moment Similarity Grid, Data Set 1.



Appendix K.7 Product Moment Similar Document Clusters, Data Set 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																							■												
2															■																																			
3															■																																			
4															■																																			
5															■																																			
6															■																																			
7		■																																																
8															■																																			
9															■																																			
10															■																																			
11															■																																			
12																																							■												
13																																							■												
14																																							■												
15			■																																															
16																																							■												
17															■																																			
18															■																																			
19																																							■												
20															■																																			
21																																							■												
22																																							■												
23																																							■												
24																																							■												
25																																										■									
26															■																																			
27																																							■												
28															■																																			
29												■																																						
30															■																																			
31															■																																			
32																■																																		
33															■																																			
34															■																																			
35																																							■												
36															■																																			
37															■																																			
38															■																																			
39															■																																			
40															■																																			
41												■																																						
42																																																		
43																																							■												
44															■																																			
45															■																																			
46																																							■												
47															■																																			
48															■																																			
49															■																																			
50															■																																			

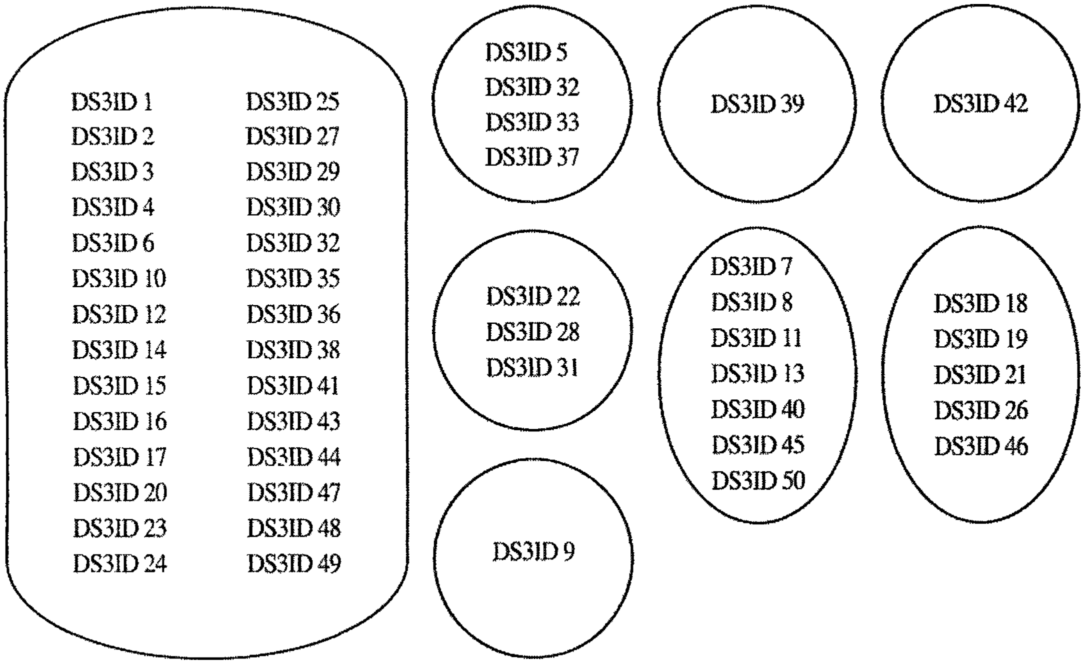
Appendix K.8 Product Moment Similarity Grid, Data Set 2.



Appendix K.9 Product Moment Similar Document Clusters, Data Set 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1								■																																										
2								■																																										
3								■																																										
4								■																																										
5																					■																													
6								■																																										
7																																				■														
8																																				■														
9																																	■																	
10								■																																										
11																																				■														
12								■																																										
13																																				■														
14								■																																										
15								■																																										
16								■																																										
17								■																																										
18		■																																																
19		■																																																
20								■																																										
21		■																																																
22																																			■															
23								■																																										
24								■																																										
25								■																																										
26		■																																																
27								■																																										
28																																			■															
29								■																																										
30								■																																										
31																																			■															
32								■																																										
33																					■																													
34																					■																													
35								■																																										
36								■																																										
37																					■																													
38								■																																										
39																		■																																
40																																				■														
41								■																																										
42																																■																		
43								■																																										
44								■																																										
45																																				■														
46		■																																																
47								■																																										
48								■																																										
49								■																																										
50																																				■														

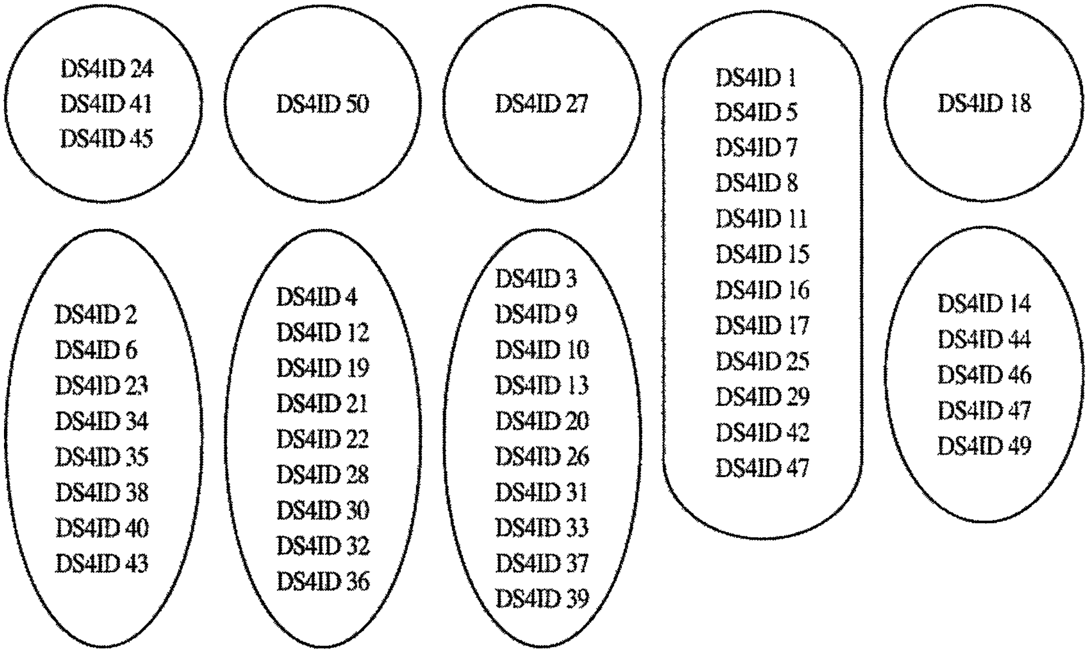
Appendix K.10 Product Moment Similarity Grid, Data Set 3.



Appendix K.11 Product Moment Similar Document Clusters, Data Set 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1																																															■			
2																					■																													
3																																						■												
4																																		■	■															
5																																															■			
6																					■																													
7																																															■			
8																																															■			
9																																						■												
10																																						■												
11																																															■			
12																																		■	■															
13																																						■												
14																		■																																
15																																															■			
16																																															■			
17																																															■			
18																																									■									
19																																		■	■															
20																																						■												
21																																		■	■															
22																																		■	■															
23																					■																													
24																																														■				
25																																															■			
26																																						■												
27									■																																									
28																																		■	■															
29																																															■			
30																																		■	■															
31																																						■												
32																																		■	■															
33																																						■												
34																					■																													
35																					■																													
36																																		■	■															
37																																						■												
38																					■																													
39																																						■												
40																					■																													
41																																														■				
42																																															■			
43																					■																													
44																		■																																
45																																														■				
46																		■																																
47																		■																													■			
48																																																		
49																		■																																
50																										■																								

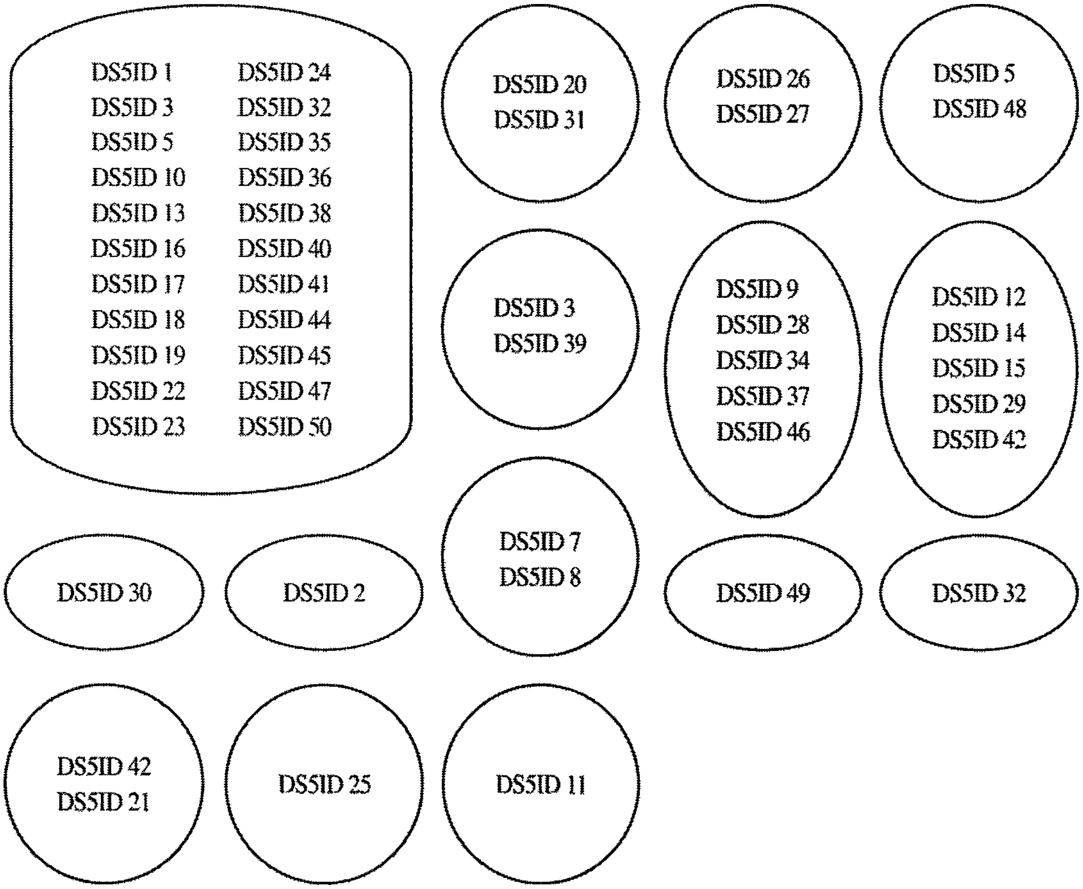
Appendix K.12 Product Moment Similarity Grid, Data Set 4.



Appendix K.13 Product Moment Similar Document Clusters, Data Set 4.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1											■																																							
2																											■																							
3											■															■																								
4																																																		
5											■																																				■			
6																																																		
7						■																																												
8						■																																												
9			■																																															
10											■																																							
11																		■																																
12																																					■													
13											■																																							
14																																					■													
15																																					■													
16											■																																							
17											■																																							
18											■																																							
19											■																																							
20														■																																				
21																																																■		
22											■																																							
23											■																																							
24											■																																							
25																							■																											
26																																							■											
27																																							■											
28			■																																															
29																																					■													
30							■																																											
31														■																																				
32											■																																							
33																																												■						
34			■																																															
35											■																																							
36											■																																							
37			■																																															
38											■																																							
39																										■																								
40											■																																							
41											■																																							
42																																					■													
43																																																■		
44											■																																							
45											■																																							
46			■																																															
47											■																																							
48																																															■			
49					■																																													
50											■																																							

Appendix K.14 Product Moment Similarity Grid, Data Set 5.



Appendix K.15 Product Moment Similar Document Clusters, Data Set 5.

REFERENCES

[1] Brill, Eric. "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging." *Eric Brill's Electronic Papers Page.* 1997. Johns Hopkins University, Baltimore, MD. 1 Feb. 2007. <http://www.cs.jhu.edu/~brill/acadpubs.htm>.

[2] Brill, Eric. "Rule-Based Tagger (Link to Source Code)." *Eric Brill's Home Page.* 1994. Johns Hopkins University, Baltimore, MD. 1 Feb. 2007. <http://www.cs.jhu.edu/~brill>.

[3] Brill, Eric. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging." *Computational Linguinstics* 21.4 (1995): 543-565.

[4] Fellbaum, Christianne, ed. *WordNet: An Electronic Lexical Database.* Cambridge, MA: The MIT Press, 2000.

[5] Hafer, Margaret A., Weiss, Stephen, F. "Word Segmentation by Letter Successor Varities." *Information Storage Retrieval* 10 (1974): 371-385.

[6] Lewis, David. D. "Reuters-21578, Distribution 1.0 Test Collection." *Reuters-21578.* Sept. 1997. Chicago, IL. 10 May 2007 <http://www.daviddlewis.com/resources/testcollections/reuters21578>.

[7] Marcus, Mitchell, et al. "The Penn Treebank: Annotating Predicate Argument Structure." *The Penn Treebank Project.* Feb. 1999. University of Pennsylvania, Philadelphia, PA. 21 Jan. 2007 <http://www.cis.upenn.edu/~treebank>.

[8] Mihalcea, Rada. "Turning WordNet into an Information Retrieval Resource: Systematic Polysemy and Conversion to Hierarchical Codes." *International Journal of Pattern Recognition and Artificial Intelligence* 17.5 (Aug 2003): 689-704.

[9] Miller, George, A., et al. "Introduction to WordNet: An On-line Lexical Database." *WordNet.* Aug. 1993. Princeton University, Princeton, NJ. 21 Jan. 2007 <http://wordnet.princeton.edu>.

[10] Nosofsky, Robert M. “Similarity, Frequency, and Category Representations.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.1 (1988): 54-65.

[11] Tversky, Amos. “Features of Similarity.” *Psychological Review* 84.4 (July 1977): 327-352.

[12] Tversky, Amos, and Gati, Itamar. “Studies of Similarity.” *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1978: 79-98.

VITA

Kimberly Adams was born in Sarasota, Florida on February 17, 1970, the daughter of Margaret and Ronnie Ball. She graduated from Venice High School in Venice, Florida, in 1988. She received the degree of Bachelor of Arts, Magna cum Laude, from Texas State University-San Marcos, previously Southwest Texas State University, in December, 1995. After eight years of working in advertising as an art director, she returned to Texas State University to study Computer Science, in June of 2003. In May of 2005, she received a Certificate in Computer Science from Texas State University. She entered The Graduate College at Texas State University in August, 2005. In April, 2007, she was recognized for Graduate Academic Excellence by the Department of Computer Science at Texas State University.

Permanent Address: 149 Steele
Kyle, Texas 78640

This thesis was typed by Kim Adams.