Using Long Runs as Predictors of Semantic Coherence in a Partial Document Retrieval System

Hyopil Shin
Computing Research Laboratory, NMSU
PO Box 30001
Las Cruces, NM, 88003
hshin@crl.nmsu.edu

Jerrold F. Stach
Computer Science Telecommunications, UMKC
5100 Rockhill Road
Kansas City, MO, 64110
stach@cstp.umkc.edu

Abstract

We propose a method for dealing with semantic complexities occurring in information retrieval systems on the basis of linguistic observations. Our method follows from an analysis indicating that long runs of content words appear in a stopped document cluster, and our observation that these long runs predominately originate from the prepositional phrase and subject complement positions and as such, may be useful predictors of semantic coherence. From this linguistic basis, we test three statistical hypotheses over a small collection of documents from different genre. By coordinating thesaurus semantic categories (SEMCATs) of the long run words to the semantic categories of paragraphs, we conclude that for paragraphs containing both long runs and short runs, the SEMCAT weight of long runs of content words is a strong predictor of the semantic coherence of the paragraph.

Introduction

One of the fundamental deficiencies of current information retrieval methods is that the words searchers use to construct terms often are not the same as those by which the searched information has been indexed. There are two components to this problem, synonymy and polysemy (Deerwester et. al., 1990). By definition of polysemy, a document containing the search terms or indexed with the search terms is not necessarily relevant. Polysemy contributes heavily to poor precision. Attempts to deal with the synonymy problem have relied on intellectual or automatic term expansion, or the construction of a thesaurus.

Also the ambiguity of natural language causes semantic complexities that result in poor precision. Since queries are mostly formulated as words or phrases in a language, and the expressions of a language are ambiguous in many cases, the system must have ways to disambiguate the query.

In order to resolve semantic complexities in information retrieval systems, we designed a method to incorporate semantic information into current IR systems. Our method (1) adopts widely used Semantic Information or Categories, (2) calculates Semantic Weight based on probability, and (3) (for the purpose of verifying the method) performs partial text retrieval based upon Semantic Weight or Coherence to overcome cognitive overload of the human agent. We make two basic assumptions: 1. Matching search terms to semantic categories should improve retrieval precision. 2. Long runs of content words have a linguistic basis for Semantic Weight and can also be verified statistically.

1 A Brief Overview of Previous Approaches

There have been several attempts to deal with complexity using semantic information. These methods are hampered by the lack of dictionaries containing proper semantic categories for classifying text. Semantic methods designed by Boyd et. al. (1994) and Wendlandt et. al. (1991) demonstrate only simple examples and are restricted to small numbers of words. In order to overcome this

deficiency, we propose to incorporate the structural information of the thesaurus, semantic categories (SEMCATs). However, we must also incorporate semantic categories into current IR systems in a compatible manner. The problem we deal with is partial text retrieval when all the terms of the traditional vector equations are not known. This is the case when retrieval is associated with a near real time filter, or when the size or number of documents in a corpus is unknown. In such cases we can retrieve only partial text, a paragraph or page. But since there is no document wide or corpus wide statistics, it is difficult to judge whether or not the text fragment is relevant. The method we employ in this paper identifies semantic "hot spots" in partial text. These "hot spots" are loci of semantic coherence in a paragraph of text. Such paragraphs are likely to convey the central ideas of the document.

We also deal with the computational aspects of partial text retrieval. We use a simple stop/stem method to expose long runs of context words that are evaluated relative to the search terms. Our goal is not to retrieve a highly relevant sentence, but rather to retrieve a portion of text that is semantically coherent with respect to the search terms. This locale can be returned to the searcher for evaluation and if it is relevant, the search terms can be refined. This approach is compatible with Latent Semantic Indexing (LSI) for partial text retrieval when the terms of the vector space are not known. LSI is based on a vector space information retrieval method that has demonstrated improved performance over the traditional vector space techniques. So when incorporating semantic information, it is necessary to adopt existing mathematical methods including probabilistic methods and statistical methods.

2 Theoretical Background

2.1 Long Runs

Partial Information Retrieval has to with detection of main ideas. Main ideas are topic sentences that have central meaning to the text. Our method of detecting main idea paragraphs extends from Jang (1997) who observed that after stemming and stopping a document, long runs of content words cluster. Content word runs are a sequence of content words with a function word(s) prefix and suffix. These runs can be weighted for density in a stopped document and vector processed. We observed that these long content word runs generally originate from the prepositional phrase and subject complement positions, providing a linguistic basis for a dense neighbourhood of long runs of content words signalling a semantic locus of the writing. We suppose that these neighbourhoods may contain main ideas of the text. In order to verify this, we designed a methodology to incorporate semantic features into information retrieval and examined long runs of content words as a semantic predictor.

We examined all the long runs of the Jang (1997) collection and discovered most of them originate from the prepositional phrase and subject complement positions. According to Halliday (1985), a preposition is explained as a minor verb. It functions as a minor Predicator having a nominal group as its complement. Thus the internal structure of 'across the lake' is like that of 'crossing the lake', with a non-finite verb as Predicator (thus our choice of ≥ 3 words as a long run). When we interpret the preposition as a "minor Predicator" and "minor Process", we are interpreting the prepositional phrase as a kind of minor clause. That is, prepositional phrases function as a clause and their role is predication.

Traditionally, predication is what a statement says about its subject. A named predication corresponds to an externally defined function, namely what the speaker intends to say his or her subject, i.e. their referent. If long runs largely appear in predication positions, it would suggest that the speaker is saying something important and the longer runs of content words would signal a locus of the speaker's intention.

Extending from the statistical analysis of Jang (1997) and our observations of those long runs in the collection, we give a basic assumption of our study:

> Long runs of content words contain significant semantic information that a speaker wants to express and focus, and thus are semantic indicators or loci or main ideas.

In this paper, we examine the SEMCAT values of long and short runs, extracted from a random document of the collection in Jang (1997), to determine if the SEMCAT weights of long runs of content words are semantic predictors.

2.2 SEMCATs

We adopted Roget's Thesaurus for our basic semantic categories (SEMCATs). We extracted the semantic categories from the online Thesaurus for convenience. We employ the 39 intermediate categories as basic semantic information, since the 6 main categories are too general, and the many sub-categories are too narrow to be taken into account. We refer to these 39 categories as SEMCATs.

Table 1: Semantic Categories (SEMCATs)

	Abbreviation	***Full Description***
1	AFIG	Affection in General
2	ANT	Antagonism
3	CAU	Causation
4	CHN	Change
5	COIV	Conditional Intersocial Volition
6	CRTH	Creative Thought
7	DIM	Dimensions
8	EXIS	Existence
9	EXOT	Extension of Thought
10	FORM	Form
11	GINV	General Inter social Volition
12	INOM	Inorganic Matter
13	MECO	Means of Communication
14	MFRE	Materials for Reasoning
15	MIG	Matter in general
16	MOAF	Moral Affections
17	MOCO	Modes of Communication
18	MOT	Motion
19	NOIC	Nature of Ideas Communicated
20	NUM	Number
21	OPIG	Operations of Intelligence In General
22	ORD	Order
23	ORGM	Organic Matter
24	PEAF	Personal Affections
25	PORE	Possessive Relations
26	PRCO	Precursory Conditions and Operations
27	PRVO	Prospective Volition
28	QUAN	Quantity
29	REAF	Religious Affections
30	RELN	Relation
31	REOR	Reasoning Organization
32	REPR	Reasoning Process
33	ROVO	Result of Voluntary Action
34	SIG	Space in General
35	SIVO	Special Inter social Volition
36	SYAF	Sympathetic Affections
37	TIME	Time
38	VOAC	Voluntary Action
39	VOIG	Volition in General

2.3 Indexing Space and Stop Lists

Many of the most frequently occurring words in English, such as "the," "of," "and," "to," etc. are non-discriminators with respect to information filtering. Since many of these function words make up a large fraction of the text of most documents, their early elimination in the indexing process speeds processing, saves significant amounts of index space and does not compromise the filtering process. In the Brown Corpus, the frequency of stop words is 551,057 out of 1,013,644 total words. Function words therefore account for about 54.5% of the tokens in a document.

The Brown Corpus is useful in text retrieval because it is small and efficiently exposes content word runs. Furthermore, minimizing the document token size is very important in NLP-based methods, because NLP-based methods usually need much larger indexing spaces than statistical-based methods due to processes for tagging and parsing.

3 Experimental Basis

In order to verify that long runs contribute to resolve semantic complexities and can be used as predictors of semantic intent, we employed a probabilistic, vector processing methodology.

3.1 Revised Probability and Vector Processing

In order to understand the calculation of SEMCATs, it is helpful to look at the structure

of a preprocessed document. One document "Barbie" in the Jang (1997) collection has a total of 1,468 words comprised of 755 content words and 713 function words. The document has 17 paragraphs. Filtering out function words using the Brown Corpus exposed the runs of content words as shown in Figure 1.

Figure1: Preprocessed Text Document

BARBIE * * * * **FAVORITE COMPANION DETRACTORS LOVE** * * * PLASTIC PERFECTION * * FASHION DOLL * * IMPOSSIBLE FIGURE * LONG * * * POPULAR GIRL * MATTEL * WORLD * TOYMAKER * PRODUCTS RANGE * FISHER PRICE INFANT * SALES * * * TALL MANNEQUIN * BARBIE * * AGE * * * **BEST SELLING GIRLS BRAND** * * POISED * STRUT * * * CHANGE * * MALE DOMINATED WORLD * MULTIMEDIA SOFTWARE * VIDEO GAMES

In Figure 1, asterisks occupy positions where function words were filtered out. The bold type indicates the location of the longest runs of content words. The run length distribution of Figure 1 is shown below:

Table 2: Distribution of Content Run Lengths in a sample Document

Run Length	*Frequency*
1	11
2	8
3	2
4	2

The traditional vector processing model requires the following set of terms:

- (df) the number of documents in the collection that each word occurs in
- (idf) the inverse document frequency of each word determined by $\log_{10}(N/df)$ where N is the total number of documents. If a word appears in a query but not in a document, its idf is undefined.
- The category probability of each query word.

Wendlandt (1991) points out that it is useful to retrieve a set of documents based upon key words only, and then considers only those documents for semantic category and attribute analysis. Wendlandt (1991) appends the *s* category weights to the *t* term weights of each document vector D_i and the Query vector Q.

Since our basic query unit is a paragraph, document frequency (df) and inverse document frequency (idf) have to be redefined. As we pointed out in Section 1, all terms are not known in partial text retrieval. Further, our approach is based on semantic weight rather than word frequency. Therefore any frequency based measures defined by Boyd et al. (1994) and Wendlandt (1991) need to be built from the probabilities of individual semantic categories. Those modifications are described below. As a simplifying assumption, we assume SEMCATs have a uniform probability distribution with regard to a word.

3.2 Calculating SEMCATs

Our first task in computing SEMCAT values was to create a SEMCAT dictionary for our method. We extracted SEMCATs for every word from the World Wide Web version of Roget's thesaurus. SEMCATs give probabilities of a word corresponding to a semantic category. The content word run 'favorite companion detractors love' is of length 4. Each word of the run maps to at least one SEMCAT. The word 'favorite' maps to categories 'PEAF and SYAF'. 'companion' maps to categories 'ANT, MECO, NUM, ORD, ORGM, PEAF, PRVO, QUAN, and SYAF'. 'detractor' maps to 'MOAF'. 'love' maps to 'AFIG, ANT, MECO, MOAF, MOCO, ORGM, PEAF, PORE, PRVO, SYAF, and VOIG'. We treat the long runs as a semantic core from which to calculate SEMCAT values. SEMCAT weights are calculated based on the following equations.

Eq.1 P_{jk}(Probability) - The likelihood of SEMCAT S_j occurring due to the K^{th} trigger. For example, assuming a uniform probability distribution, the category PEAF triggered by the word favorite above, has the following probability:

$$P_{PEAF,\,favorite} = 0.5(1/2)$$

Eq.2 Sw_j (SEMCAT Weights in Long runs) is the sum of each SEMCAT(j) weight of long runs based on their probabilities. In the above example, the long run

'favorite companion detractors love,' the SEMCAT 'MOAF' has Sw_{MOAF} : (detractor(1) + love(.09)) = 1.09. We can write;

$$Sw_j = \sum_{i=1}^{r} P_{ij}$$

Eq.3 edw_j (Expected data weights in a paragraph) - Given a set of N content words (data) in a paragraph, the expected weight of the SEMCATs of long runs in a paragraph is:

$$edw_j = \sum_{i=1}^{N} P_{ij}$$

Eq.4 idw_j (Inverse data weights in a paragraph) - The inverse data weight of SEMCATs of long runs for a set of N content words in a paragraph is

$$idw_j = \log_{10}\left(\frac{N}{edw_j}\right)$$

Eq.5 Weight(W_j) - The weight of SEMCAT S_j in a paragraph is

$$W_j = Sw_j \times idw_j$$

Eq.6 Relevance Weights (Semantic Coherence)

$$W = \sum_{i=1}^{r} W_{ij}$$

Our method performs the following steps:

1. calculate the SEMCAT weight of each long content word run in every paragraph (Sw)
2. calculate the expected data weight of each paragraph (edw)
3. calculate the inverse expected data weight of each paragraph (idw)
4. calculate the actual weight of each paragraph (Sw×idw)
5. calculate coherence weights (total relevance) by summing the weights of (Sw×idw).

In every paragraph, extraction of SEMCATs from long runs is done first. The next step is finding the same SEMCATs of long runs through every word in a paragraph (expected data weight), then calculate idw, and finally Sw×idw. The final, total relevance weights are an accumulation of all weights of SEMCATs of content words in a paragraph. Total relevance tells how many SEMCATs of the Query's long runs appear in a paragraph. Higher values imply that the paragraph is relevant to the long runs of the Query.

The following is a program output for calculating SEMCAT weights for an arbitrary long run: "SEVEN INTERACTIVE PRODUCTS LED"

SEMCAT: EXOT Sw : 1.00 edw : 1.99 idw : 1.44 Sw×idw : 1.44
SEMCAT: GINV Sw : 0.33 edw : 1.62 idw : 1.53 Sw×idw : 0.51
SEMCAT: MOT Sw : 0.20 edw : 0.71 idw : 1.89 Sw×idw : 0.38
SEMCAT: NUM Sw : 0.20 edw : 1.76 idw : 1.49 Sw×idw : 0.30
SEMCAT: ORGM Sw : 0.20 edw : 1.67 idw : 1.52 Sw×idw : 0.30
SEMCAT: PEAF Sw : 0.53 edw : 1.50 idw : 1.56 Sw×idw : 0.83
SEMCAT: REAF Sw : 0.20 edw : 0.20 idw : 2.44 Sw×idw : 0.49
SEMCAT: SYAF Sw : 0.33 edw : 1.19 idw : 1.66 Sw×idw : 0.55

Total (Sw×idw) : 4.79

4 Experimental Results

The goal of employing probability and vector processing is to prove the linguistic basis that long runs of content words can be used as predictors of semantic intent But we also want to exploit the computational advantage of removing the function words from the document, which reduces the number of tokens processed by about 50% and thus reduces vector space and probability computations. If it is true that long runs of content words are predictors of semantic coherence, we can further reduce the complexity of vector computations: (1) by eliminating those paragraphs without long runs from consideration, (2) within remaining paragraphs with long runs, computing and summing the semantic coherence of the longest runs only, (3) ranking the eligible paragraphs for retrieval based upon their semantic weights relative to the query.

Jang (1997) established that the distribution of long runs of content words and short runs of content words in a collection of paragraphs are drawn from different populations. This implies

that either long runs or short runs are predictors, but since all paragraphs contain short runs, i.e. a single content word separated by function words, only long runs can be useful predictors. Furthermore, only long runs as we define them can be used as predictors because short runs are insufficient to construct the language constructs for prepositional phrase and subject complement positions. If short runs were discriminators, the linguistic assumption of this research would be violated. The statistical analysis of Jang (1997) does not indicate this to be the case.

To proceed in establishing the viability of our approach, we proposed the following experimental hypotheses:

(H1) The SEMCAT weights for long runs of content words are statistically greater than weights for short runs of content words. Since each content word can map to multiple SEMCATs, we cannot assume that the semantic weight of a long run is a function of its length. The semantic coherence of long runs should be a more granular discriminator.

(H2) For paragraphs containing long runs and short runs, the distribution of long run SEMCAT weights is statistically different from the distribution of short run SEMCAT weights.

(H3) There is a positive correlation between the sum of long run SEMCAT weights and the semantic coherence of a paragraph, the total paragraph SEMCAT weight.

A detailed description of these experiments and their outcome are described in Shin (1997, 1999). The results of the experiments and the implications of those results relative to the method we propose are discussed below. Table 3 gives the SEMCAT weights for seventeen paragraphs randomly chosen from one document in the collection of Jang (1997).

Table 3: SEMCAT Weights of 17 Paragraphs Chosen Randomly From a Collection

Paragraph	Short Runs Weight	Long Runs Weight
1	29.84	18.60
2	31.29	12.81
3	23.29	4.25
4	23.94	11.63
5	34.63	35.00
6	22.85	03.32
7	21.74	00.00
8	35.84	15.94
9	30.15	00.00
10	13.40	00.00
11	23.01	07.82
12	31.69	04.79
13	36.54	00.00
14	17.91	10.55
15	19.70	05.83
16	17.11	00.00
17	31.86	00.00

The data was evaluated using a standard two way F test and analysis of variance table with $\alpha = .05$. The analysis of variance table for the paragraphs in Table 3 is shown in Table 4.

Table 4: Analysis of Variance for Table 2 Data

Variation	Degrees of Freedom	Mean Square	F
Between Treatments $V_R = 2904.51$	1	2904.51	68.56
Between Blocks $V_C = 1502.83$	16	93.92	2.21
Residual or Random $V_E = 677.77$	16	42.36	
Total $V = 5085.11$	33		

At the .05 significance level, $F_{\alpha = .05} = 4.49$ for 1,16 degrees of freedom. Since $68.56 > 4.49$ we reject the assertion that column means (run weights) are equal in Table 2. Long run and short run weights come from different populations. We accept H1.

For the between paragraph treatment, the row means (paragraph weights) have an F value of 2.21. At the .05 significance level, $F_{\alpha = .05} = 2.28$ for 16,16 degrees of freedom. Since $2.21 < 2.28$ we cannot reject the assertion that there is no significant difference in SEMCAT weights between paragraphs. That is, paragraph weights do not appear to be taken from different populations, as do the long run and short run weight distributions. Thus, the semantic weight

of the content words in a paragraph cannot be used to predict the semantic weight of the paragraph. We therefore proceed to examine H2.

Notice that two paragraphs in Table 2 are without long runs. We need to repeat the analysis of variance for only those paragraphs with long runs to see if long runs are discriminators. Table 5 summarizes those paragraphs.

Table 5: SEMCAT weights of 11 paragraphs containing long runs and short runs

Paragraph	Short Runs Weight	Long Runs Weight
1	29.84	18.60
2	31.29	12.81
3	23.29	4.25
4	23.94	11.63
5	34.63	35.00
6	22.85	03.32
8	35.84	15.94
11	23.01	07.82
12	31.69	04.79
14	17.91	10.55
15	19.70	05.83

This data was evaluated using a standard two way F test and analysis of variance with $\alpha = .05$. The analysis of variance table for the paragraphs in Table 5 follows.

Table 6: Analysis of Variance for Table 5 Data

Variation	Degrees of Freedom	Mean Square	F
Between Treatments $V_R = 1430.98$	1	1430.98	291.44
Between Blocks $V_C = 944.08$	10	94.40	19.22
Residual or Random $V_E = 49.19$	10	4.91	
Total $V = 2424.26$	21		

At the .05 significance level, $F_{\alpha = .05} = 4.10$ for 2,10 degrees of freedom. $4.10 < 291.44$. At the .05 significance level, $F_{\alpha = .05} = 2.98$ for 10,10 degrees of freedom. $2.98 < 19.22$. For paragraphs in a collection containing both long and short runs, the SEMCAT weights of the long runs and short runs are drawn from different distributions. We accept H2.

For paragraphs containing long runs and short runs, the distributions of long run SEMCAT weights is different from the distribution of short run SEMCAT weights. We know from the linguistic basis for long runs that short runs cannot be used as predictors. We therefore proceed to examine the Pearson correlation between the long run SEMCAT weights and paragraph SEMCAT weights for those paragraphs with both long and short content word runs.

Table 7: Correlation of Long Run SEMCAT Weights to Paragraph SEMCAT Weight

Paragraph	Long Runs Semantic Weight	Paragraph Semantic Weight
1	18.60	48.44
2	12.81	44.10
3	4.25	27.54
4	11.63	35.57
5	35.00	69.63
6	03.32	26.17
8	15.94	51.78
11	07.82	30.83
12	04.79	31.69
14	10.55	28.46
15	05.83	25.53

The weights in Table have a positive Pearson Product Correlation coefficient of .952. We therefore accept H3. There is a positive correlation between the sum of long run SEMCAT weights and the semantic coherence of a paragraph, the total paragraph SEMCAT weight.

5. Conclusion

This research tested three statistical hypotheses extending from two observations: (1) Jang (1997) observed the clustering of long runs of content words and established the distribution of long run lengths and short run lengths are drawn from different populations, (2) our observation that these long runs of content words originate from the prepositional phrase and subject complement positions. According to Halliday (1985) those grammar structures function as

minor predication and as such are loci of semantic intent or coherence. In order to facilitate the use of long runs as predictors, we modified the traditional measures of Boyd et al. (1994), Wendlandt (1991) to accommodate semantic categories and partial text retrieval. The revised metrics and the computational method we propose were used in the statistical experiments presented above. The main findings of this work are

1. the distribution semantic coherence (SEMCAT weights) of long runs is not statistically greater than that of short runs,
2. for paragraphs containing both long runs and short runs, the SEMCAT weight distributions are drawn from different populations
3. there is a positive correlation between the sum of long run SEMCAT weights and the total SEMCAT weight of the paragraph (its semantic coherence).

Significant additional work is required to validate these preliminary results. The collection employed in Jang (1997) is not a standard Corpus so we have no way to test precision and relevance of the proposed method. The results of the proposed method are subject to the accuracy of the stop lists and filtering function.

Nonetheless, we feel the approach proposed has potential to improve performance through reduced token processing and increased relevance through consideration of semantic coherence of long runs. Significantly, our approach does not require knowledge of the collection.

References

Boyd R., Driscoll J, and Syu I. (1994) *Incorporating Semantics Within a Connectionist Model and a Vector Processing Model.* In Proceedings of the TREC-2, NIST.

Deerwester S., Furnas G., Landauer T., and Harshman R. (1990) Indexing by Latent Semantic Anaysis. Journal of the American Society of Information Science 41-6.

Halliday M.A.K. (1985) An Introduction to Functional Grammar. Edward Arnold, London.

Jang S. (1997) Extracting Context from Unstructured Text Documents by Content Word Density. M.S. Thesis, University of Missouri-Kansas City.

Moffat A., Davis R., Wilkinson, R., and Zobel J. (1994) Retrieval of Partial Documents. In Proceedings of TREC-2.

Shin H. (1997) Incorporating Semantic Categories (SEMCATs) into a Partial Information Retrieval System. M.S. Thesis, University of Missouri-Kansas City.

Shin H., Stach J. (1999) Incorporating Probabilistic Semantic Categories (SEMCATs) Into Vector Space Techniques for Partial Document Retrieval. Journal of Computer Science and Information Management, vol. 2, No. 4, December 1999, to appear.

Wendlandt E. and Driscoll R. (1991) Incorporating a semantic analysis into a document retrieval strategy. CACM 31, pp. 54-48.