

# Semantic Information Evolution

Hemant M. Joshi and Coskun Bayrak  
Applied Science Department  
University of Arkansas at Little Rock  
{hmjoshi | cxbayrak@ualr.edu}

## ABSTRACT

A librarian at Yale University, Mr. Rogers Rutherford says, "We're drowning in information and starving for knowledge" [9]. These quotes are very appropriate keeping in mind the rate at which World Wide Web is gathering information. It is essential to be able to make sense out of these huge piles of information. It is important to look at information from the point of view of its relevance, correlation and also from the point of view of its validity with respect to time. Semantic information will facilitate the correlation between two given concepts. The focus of the following paper is on some of the inherent intrinsic properties of information. The task of identifying any information with right context is very important. This paper deals with the concept of reference domains, which represents information as it evolves in our knowledge base.

## BACKGROUND

Sherlock Holmes and Dr. Watson go on a camping trip, set up their tent, and fall asleep. Some hours later, Holmes wakes his faithful friend. "Watson, look up at the sky and tell me what you see." Watson replies, "I see millions of stars." "What does that tell you?" Watson ponders for a minute.

"Astronomically speaking, it tells me that there are millions of galaxies and potentially billions of planets. Astrologically, it tells me that Saturn is in Leo. Time wise, it appears to be approximately a quarter past three.

Theologically, it's evident the Lord is all-powerful and we are small and insignificant. Meteorologically, it seems we will have a beautiful day tomorrow. What does it tell you?" Holmes is silent for a moment, and then speaks. "Someone has stolen our tent"

*The story is hilarious. What it highlights is how easy it is to complicate issues and then miss the most obvious things. Too much knowledge is no help unless it is supported by the wisdom to guide it to an effective conclusion.*

More efficient searching for information has been a goal for the research community since the frequent use of the World Wide Web to find every possible answer needed. It is desired to look for information encompassing all possible areas for a given concept. The scope for information retrieval should not be confined to a particular field. This paper discusses the human-like cognitive abilities to retrieve information. Information is correlated with what is known and processed with respect to the given context. Concepts are more meaningful when they are expressed with the right context.

## PROBLEM DEFINITION

Text has been the most prevalent method of representing information. Thus text pattern matching is the basis of information retrieval. Even though some progress has been made with techniques like latent semantic information and probabilistic latent semantic information, it is needed to look at information along with its context to make more intuitive judgment. Some of the techniques discussed in this paper are of utmost importance from the point of view of information representation. Information, if represented in the right context, makes more sense than text based searching.

Most of the top search engines [1-6] do a better job at matching text to the query. There is, however, limited semantics involved in this process. Same word might mean differently for different people. An Indonesian knows Java as one of the islands in

Indonesia, south of Borneo. A computer programmer knows Java as a programming language. Some coffee enthusiast would like to know about Java coffee. Most of the search engines including the popular ones do get the job done but fail to interpret and understand semantics of the same word java and its importance to various people. Google [2], [8] arguably the best search engine currently has pile of 3,307,998,701 web pages in its database and search term Java results in 56600000 hits with none of the first 100 results about Java coffee or java islands. Google's highly acclaimed Page Rank technique along with some of the best search engine practices fails to take into account semantics or different meaning of the word Java. Effective representation is important to retrieve information by association, correlation, relevance and semantics.

### PROPERTIES OF INFORMATION

Information in all its forms has some properties associated with it. These properties are inherent to the nature of information and do apply to all types of information.

1. **Temporal:** Information in its any form has temporal information associated with it. These temporal parameters talk about the validity of the information "from" and "to" date. Information is not created the day it is put on the web page but the day it is associated with. Some of the information may have unlimited validity [mostly facts] but still has temporal properties strongly associated with it.
2. **Semantics:** Semantic or meaning of the information is always a language issue, as it is dependent on the language it is expressed in. This is the reason why information processing should involve great deal of Natural Language Processing [7] as well. Semantic property also refers to the natural characteristics of different media to represent information. Considering the case of textual information, Natural Language Processing suggests that the most strongly correlated information is the one expressed in one single sentence. Information should be looked at from the context it was established in.

### PREVIOUS WORK

The standard information representation and retrieval process includes parsing of web pages, removing (called as stemming [8]) unwanted or common words and remaining information is considered good to be indexed. This information is then ranked and indexed according to various algorithms. One of the most common and effective algorithm uses Term Frequency (TF) [14] and Inverse Document Frequency (IDF) [13] to determine weight of each term. When searched for the given term, documents that match certain threshold are listed as search results. The results produced by these systems are satisfactory but not intuitive from the point of view of context of information. Traditional lexical matching is more of pattern matching technique without any semantic involvement.

Latent semantic analysis (LSA) is the first step in the direction of semantic analysis of document space. There are known problems of mere lexical matching techniques and some of these are very well addressed in LSA technique. LSA expresses term by document matrix for the given document space and then tries reducing the dimensionality by employing Singular Value Decomposition (SVD) technique. The sparse matrix of term frequencies in the given document space is resolved and mapped to a semantic space where the query vector can be compared against each document to find cosine similarity. Performance is greatly improved with this SVD technique compared to original term document matrix [16].

Probabilistic LSA technique introduced by Thomas Hoffman [17] is a variant of LSA, which by forming aspect model represents data in latent probability space. Probabilistic LSA yields better results than traditional LSA model.

According to Thomas Landauer et. al [18], LSA even though better technique than lexical matching, lacks human cognitive abilities to intuitively understand correlations. It fails to take into account contextual information.

## INTRODUCTION

The concept of semantic view of information is presented in this section. Following are the definitions of a few terms that are referred to in this paper.

- **Information Domain(s) and sub-domains**  
Information domains are those primary important word(s) in the text information, which represent given concept for its context established. Sub domains are synonyms and semantically similar meanings that constitute the entire information domain. Hierarchically sub domains are part of and belong to main domain that is encountered in the document space.
- **Reference domains**  
Various domains will attribute to the representation of information present in the given document space. If a document has two words that belong to different (sub) domains, in a single sentence, their strong correlation can easily be determined. Such (sub) domains will be combined to form a reference domain. In short, reference domains span across set of domains of information.
- **Correlation and types of correlation**  
Correlation is defined in terms of co-occurrence of two given terms occurring in the same document. Presence of these terms in a single document illustrates the correlation between them.  
*Strong Correlation* is said to exist if the two terms not only exist in the same document but also in a single sentence, as represented by reference domains.  
*Weak correlation* is more common and is handled by two terms occurring in a single document, not necessarily in a single sentence.  
*No correlation* exists in two given terms if they do not occur in a same document.
- **Related Domains**  
If two (sub) domains are (weakly/strongly) correlated in terms of their semantics, then they are considered as Related Domains. Reference domains are subset of related dataset as they represent only (positively) strong correlation. The three words term, domains and sub-domains can be used interchangeably. So strong, weak or no correlation, can be expressed in terms of two (sub) domains as well.
- **Semantic Correlation among related domains**  
As mentioned above, semantic correlations can be expressed between related domains. These relationships could be of strong, weak nature or no relation may exist semantically between two non-related domains.

## INFORMATION CORRELATION AND ITS REPRESENTATION

Information correlation between two (sub) domains is represented by the qualitative correlation between the two. Consider domains  $d_i$  of elements. If there is a reference domain  $R$  such that there can be found an element of  $R$  ( $r_{iq} \in R$ ),  $r_{jq}$  such that  $r_{jq}$  is either  $d_i$  itself or is in close vicinity of  $d_i$ . Then the qualitative correlation between  $d_i$  and  $d_i$  can be defined as:

$$C_i(d_i) = \begin{cases} 1 \\ 0 \end{cases} \quad \text{--- (1)}$$

Here the value of qualitative correlation is 1 if and only if  $d_i$  can be found in close vicinity of  $r_{jq}$  i.e.  $r_{jq} - d_0 \leq d_i \leq r_{jq} + d_0$ . Here  $d_0$  is the shift interval for the depth till which references formed within reference domains can be persuaded. This factor can be standardized to be 1, which means 1 reference to outside domain from within actual reference domain is followed and is considered significant. On the other hand, value of correlation function  $C_i(d_i)$  is determined to be 0 if  $d_i$  cannot be found in given document space or in the close vicinity of reference domain element  $r_{jq}$ . The definition of  $C_i(d_i)$

here indicates that for the given reference domain  $d_i$ , all the elements in that reference domain exhibit related domain property to establish relation with  $d_i$ .

**SEMANTIC CORRELATION FUNCTION OF REFERENCE DOMAIN**

Let there be  $m$  elements that are members of reference domain represented by  $d_i$ . In this case, Semantic Correlation Function of the entire domain  $d_i$  can be calculated by considering the qualitative correlation of each member with domain  $R_j$ . Because all the elements of domain  $d_i$  have same strong correlation to the term under consideration, the Semantic Correlation Function (SCF) for a given reference domain  $d_i$  is defined as follows:

$$SCF_i = \frac{m - 1}{m}$$

where reference domain has  $m$  elements such that  $t = 1,2,3, \dots, m$ , and,  $0 < SCF_i < 1$ . Graph in Figure 1 represents relationship between number of elements in the reference domain and its SCFi.

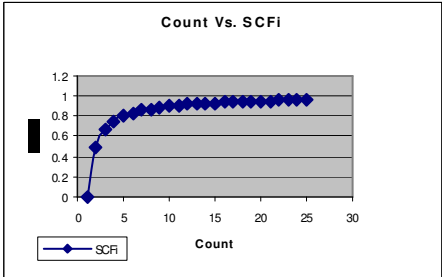


Figure 1: Relationship between number of elements in the reference domain and its SCFi

If count,  $m = 1$  i.e. only one element is member of reference domain which is under consideration, then SCFi value is 0. As the number of words in the related domain increases, the SCFi value will be close to 1. In other words, SCFi is in direct proportion to the number of elements of reference domain that can be identified as related domains.

The relationship between correlation and number of outside references can be established if the number of outside references pointing to the given reference domain are considered. For any given count, once SCFi is obtained, correlation factor can be defined as:

$$Corr_i = \log_{10} (SCF_i * (1+N)) \quad \text{--- (2)}$$

where  $N$  is the number of outside references for any given domain.

For any given number of elements in the domain, if the number of elements is kept steady and the number of outside references to that particular reference domain is increased, the value of correlation of the given domain increases linearly slightly as shown in graph of Figure 2. The increase is justified and the slight change in logarithmic function of correlation does correspond to the varying domain references that may be found in the given document space. The use of logarithmic function is suitable as there are a large number of outside domain references.

The affecting factor SCFi also plays a crucial role determining correlation factor for any given domain. The presence of 1 in the formula above ensures proper values even if there are no outside references to the given domain. Also it indicates that there is at least one reference (self referencing) to any domain at any given instance.

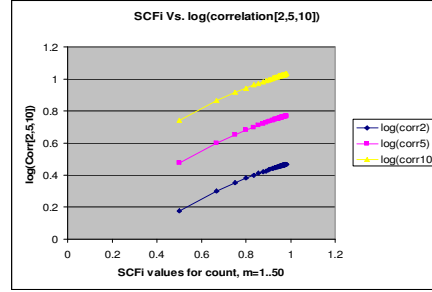


Figure 2: SCF response to number of reference domain elements

## PROPERTIES

Some important aspects of above representation can be inferred. Dynamic shift interval or depth of reference, to related domain could be adjusted for tuning of related query results. In some cases, this parameter may be omitted to prevent any depth for related domain search. In this specific case, the qualitative correlation of information domains degenerates into fuzzy logic and  $d_0$  becomes fuzzy logic interval.

Another evident interesting property is that for the same number of elements in two given domains, if number of outside references is more, the correlation value of that domain is higher. So correlation value is directly proportional to the number of outside references. Even though correlation value is directly proportional to number of outside references, it also is dependent on the SCFi value of the given domain.

These principles are applied to the document matrix method of information representation. The LSA technique is adopted to represent reference domains across the entire document space. Imagine a document space of about 500 documents, and around 2000 lines or sentences consisting of 5 word each making total of 10,000 terms. Lexical matching would identify and match any of these terms by their presence in text. No correlations can be established if the semantic aspect is not considered.

Instead of term document matrix of 10,000 x 500 size, the dimensionality is reduced by creating correlation factor matrix of reference domains and documents in the given document space. This matrix would be of considerably smaller size like 2000 x 500. The correlation factors for each reference domain correspond with respect to the given document. Singular Value Decomposition [19] to this matrix can be further applied to reduce dimension to represent data in truly semantic space. Even the query vector can be mapped to this same space. The cosine similarity between any given document vector and the query vector can be used as a measure to determine whether the particular document is similar or rather correlated to the query term. Local and global weights can be added to the equation (2) to further improve results.

An algorithm to find correlations semantically; employs LSA. The base matrix of reference domains vs. document can be split and reduced in dimensionality by using Singular Value Decomposition. The original matrix is represented by A here.

$$A = U * \Sigma * V^T$$

Here matrix A is of r reference domains and d documents, whereas Matrix U and  $V^T$  are of much smaller dimension. Both vectors U and  $V^T$  are orthogonal and vector  $\Sigma$  consists of singular values. Original matrix is formed using equation (2) in this paper. Now the first selected k values of diagonal vector  $\Sigma$  are used to reconstruct the original matrix. We form new matrix  $A_k$  such that,

$$A_k = U_k * \Sigma_k * V_k^T$$

The query vector can now be formulated. The query term defines the query vector. Querying approach looks for semantic context for the given query and forms a query reference domain. It is assumed that each reference domain has a potential to be a query result and minimum weight is given for that fact. As a result, query vector is formed of various reference domains correlation factors.

The query vector is then mapped to the semantic space of reduced dimensionality  $k$  as follows:

$$Q_k = Q^T * U_k * \Sigma_k^{-1}$$

Similarly all the document vectors can be individually mapped to semantic space as:

$$dj_k = dj_k^T * U_k * \Sigma_k^{-1}$$

where  $dj \in D$ , document space. Now that query as well as every document is mapped to the semantic space, the cosine of angle between two vectors can be defined as:

$$\text{sim}(dj, Q) = \cos \theta_j = \frac{dj_k^T * Q_k}{\|dj_k^T\|_2 * \|Q_k\|_2}$$

Once the similarity has been established, correlated domains are identified and sorted for the closeness to the query vector.

The strength of this approach is in forming reference domains of synonyms and related contexts for the given terms. Information with its relevant context is preserved and mined for information retrieval. Also significant change can be achieved in terms of original matrix dimensionality. The term by document matrix has row dimension, which is much larger than original matrix of reference domain by document.

Information can be represented as it evolves in the document space. Once the correlation circle is complete, information keeps growing, seeking more and more correlation. Not all correlated concepts are strongly related in the reality but they are significantly represented the way human mind would try and correlate them. The context in which a particular concept is absorbed makes this approach establish newer correlations.

Test case environment was set up to analyze and implement concept correlation algorithm. The performance was tested in incremental fashion linearly increasing the size of the document space. Noise documents were also introduced to simulate real world scenario. Text as well as HTML documents were parsed and matched for dictionary meanings, synonyms. Figure 3 indicates the accuracy of the system developed to implement correlation algorithm.

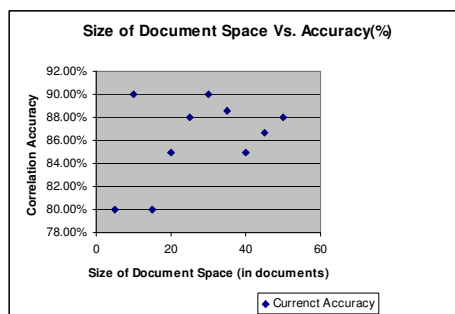


Figure 3: Accuracy in finding correlations

## FUTURE WORK AND CONCLUSION

This paper shows the semantic relations between concepts that are strongly correlated. The established relationships should help determine relevant correlated information for the given query term. This technique can be used in conjunction with some of the existing techniques like LSA to reduce space and time complexity and to provide better and relevant search results.

Future work in this area will involve larger document space to form reference domains. Also the use of encyclopedia will help determine relevance and context of given concept in broader aspect. One key area of work can be devoted to understanding sentence boundaries in different languages. It will help improve this technique in various languages. Another key issue to be investigated is about temporal properties associated with information. These properties can help us associate information with its relevance and with respect to time features as well.

## REFERENCES

- [1] AltaVista web page search engine, <http://www.altavista.com>
- [2] Google Search engine, <http://www.google.com>
- [3] Vivisimo, cluster search engine, <http://www.vivisimo.com>
- [4] Seeq search engine, <http://www.seeq.com>
- [5] Kartoo, visual categorized search engine, <http://www.kartoo.com>
- [6] Web page search engine, <http://www.webcrawler.com>
- [7] Foundations of Statistical Natural Language processing, <http://nlp.stanford.edu/fsnlp/>
- [8] Porter stemming algorithm, <http://www.tartarus.org/~martin/PorterStemmer/>
- [9] Information drowning quote, <http://www.bartleby.com/63/5/2605.html>
- [10] "Using Qualitative Hypotheses to Identify Inaccurate Data" By Qi Zhao and Toyoki Nishida, Journal of Artificial Intelligence Research 3, (1995) pp.119-145
- [11] "Towards Semantic Web Mining" By Bettina Berendt, Andreas Hotho, and Gerd Stumme, ISWC 2002, LNCS 2342, pp. 264-278, <http://citeseer.nj.nec.com/585344.html>
- [12] "The anatomy of a Large-Scale Hyper textual Web Search Engine" By Sergey Brin and Lawrence Page, <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>
- [13] "Text categorization based on weighted inverse document frequency" By Tokenobu Tokunaga and Makoto Iwayama (1994), <http://citeseer.nj.nec.com/tokunaga94text.html>
- [14] Term Frequency definition by search engine dictionary at <http://www.searchenginedictionary.com/terms-term-frequency.shtml>
- [15] "Computing Iceberg Concept Lattices with TITANIC" By Gerd Stumme et. al (2002) at <http://citeseer.ist.psu.edu/stumme02computing.html>
- [16] "Using Linear Algebra For Intelligent Information Retrieval" By Michael Berry et. al. at, Technical Report UT-CS-94-270, 1994 on the web at <http://citeseer.nj.nec.com/berry95using.html>
- [17] "Unsupervised Learning by Probabilistic Latent Semantic Analysis" By Thomas Hoffman, Machine Learning, 42, 177-196, 2001 on the web at <http://ai.uwaterloo.ca/~sjwang/cs886/hofmann.pdf>
- [18] "An introduction to Latent Semantic Analysis" By Thomas K. Landauer et.al, Discourse Processes, 25, 259-284 on the web at <http://lsa.colorado.edu/papers/dp1.LSAintro.pdf>

- [19] "Learning Human-like knowledge by Singular Value Decomposition : A progress Report" By Thomas Lanauer, Darrell Laham and Peter Foltz, Artificial intelligence on the web at <http://lsa.colorado.edu/papers/nips.pdf>
- [20] "The Theory Underlying Concept Maps and How To Construct Them" By Joseph D. Novak, Cornell University on the web at <http://cmap.coginst.uwf.edu/info/>