IADIS WWW/Internet Conference", pp 1129-1134, Madrid, Spain, 6-9 October, 2004.

"SURFING FOR KNOWLEDGE" FINDING SEMANTICALLY SIMILAR WEB CLUSTERS

David Cleary
Applied Research Labs, Ericsson Ireland.
Ericsson Software Campus, Athlone, Ireland
David.Cleary@ericsson.com

Diarmuid O' Donoghue
Department. of Computer Science, NUI Maynooth, Ireland.
Maynooth, Co. Kildare, Ireland.
diarmuid.odonoghue@may.ie

ABSTRACT

In this paper we present our technique for finding semantically similar clusters within web documents obtained from a set of queries retrieved from the Google search engine. This technique utilizes a clustering algorithm based on previous Latent Semantic Analysis (LSA) work pioneered by Deerwester. In this paper we demonstrate how by using our clustering algorithm we can resolve ambiguities prevalent in natural language such as polysemy and synonymy. Following from a detailed description of the algorithm we present our initial findings using real world Internet queries. We conclude by evaluating the merits of our clustering algorithm through comparison with results observed by human categorization.

KEYWORDS

Information Retrieval, Semantic Web, Latent Semantic Analysis.

1. INTRODUCTION

When retrieving information from a search engine, the ability to identify documents related to the meaning of a query is of utmost importance. Typically, the identified documents relate to several different interpretations of the supplied query terms, with documents related to each interpretation randomly scattered across the returned documents. One possible approach is to apply automatic knowledge filters to the information retrieval process.

The most challenging issues in web search centre on natural language ambiguity. Both web pages and search queries are expressed in natural language, and thus suffer from ambiguity. *Synonymy* (multiple lexemes with the same meaning *eg. path* and *pavement*) and *polysemy* (one lexeme with multiple meanings, *eg. cook* could refer to explorer or food preparation) are inherent ambiguities in most hypertext documents; and attempts to counteract such problems have proved difficult. The Semantic Web (Berners-Lee *et al*, 2001) initiative has adopted the ontological approach (Gruber, 1993; Deborah *et al*, 2003), where documents are associated with a vocabulary and a context in which the vocabulary is valid. This approach requires discipline when publishing content and is effective in machine-to-machine interactions. However, due to its artificial nature, it has limited usage for the informal documents that characterize most of the WWW. The reality of the Internet is that documents and terms will often be spread over several topics; thus, some topics will not be as sharply defined as others limiting the use of an ontological approach.

In this paper we attempt to find clusters of pages that have related semantic meaning. These clusters rely on extracting topics contained within a sub-set of the web using Latent Semantic Analysis (LSA) techniques pioneered by Deerwester (Deerwester *et al*, 1990). This statistical model of word usage allows processing of information into structures that take advantage of the implicit higher order associations of words within a document corpus. This technique uses *implicit* semantic information to discover information clusters, where refined queries can be used on the "*ontology-less"* web. By identifying semantically similar clusters of information we can start to create a truer understanding of information retrieved from ambiguous queries.

2. RELATED WORK

Initial work on identifying web communities (Pirolli *et al*, 1996) (Flake *et al*, 2002) focuses on modeling the web as a graph where vertices are web pages and hyperlinks are edges. The foundation of these approaches stems from Kleinbergs work on hubs and authorities (Klienberg, 1998) and the popularised PageRank (Page and Brin, 1998) by Google. Other approaches based on graph theory of note are bipartite subgraph (Kumar *et al*, 1999) identification, spreading activation energy (SAE) (Pirolli *et al*, 1996). The above approaches are all based on graph theory, thus when applying these techniques to the Internet, meaningful results are only achieved from these approaches when textual information is added.

Latent Semantic Indexing (LSI) is a statistical information retrieval method designed to overcome two common problems in information retrieval synonymy and polysemy. This technique has being shown to work well when satisfying several technical assumptions (Azar *et al*, 2001). Many mathematical models based on LSA have been derived (Dumais, 1996) to create more accurate results with later activities centering on probabilistic techniques (Papadimitriou *et al*, 1998). Work addressing link structure with user queries and web content (Achlioptas *et al*, 2001) has derived rigorous frameworks for web searching. Our work uses the richer understanding of similarity that LSA offers to identify cluster of similar information. By expanding the use of content as apposed to simple topological information we can group semantically similar web content allowing for more focused retrieval of ambiguous topics from the web.

3. IDENTIFYING WEB CLUSTERS

The first stage needed to determine clusters of information is to represent each hyperlinked page as a vector V_i. This is achieved by creating a dictionary of all terms that occur in a web document. Removing a set of commonly occurring stopwords such as prepositions and articles refines this initial raw set. The remaining words are stored with their corresponding frequency of occurrence for that given document. So as to reduce the processing requirements, a subset comprising of the most frequent words are kept. This process is repeated for all *d* documents on which the clustering algorithm is carried out.

The resulting information about each page can be encoded as a vector V_i resulting in the formation of the library $L=\{\ V_{1,}\ V_2\ V_3\ \ldots\ldots V_d\ \}$ where each document vector $\mathrm{V_i}$ is a set of frequently occurring terms associated with each document. We now construct a document term matrix *X* representing the library *L* where each row represents a word and each column represents a document. The contents of each cell in the matrix, X_{ij} corresponds to the frequency of the word *i* occurring in document *j*. Therefore the column vector X_j is the representation of document *i* over the entire concept space of terms in *L*.

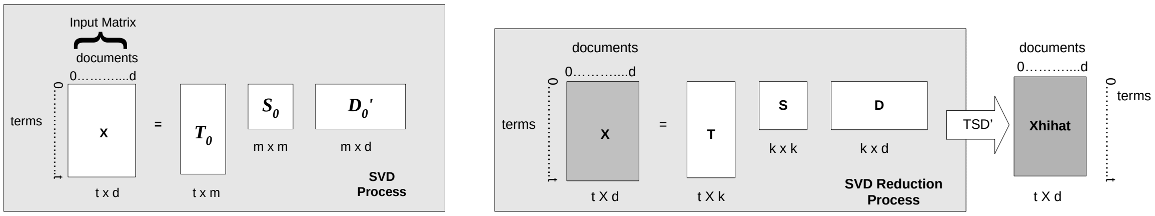

Figure 1 SVD and Dimension reduction using LSA

X is transformed using single vector decomposition (SVD) (Deerwester *et al*, 1990) into $X= T_0\ S_0\ D_0'$ as seen in figure 1. If we now order the singular values in S_0 and reduce the matrix to a square matrix of size *k* we can create a new matrix *Xhihat*, the product of the resulting matrixes. *Xhihat≈X≈TSD.* This matrix *Xhihat* is only an approximation equal to *X* with a reduced model that is the best possible least-square-fit to *X*. It is this Xhilhat matrix that we use as the basis for our cluster analysis. This transformation reduces the dimensionality by ‘projecting’ the tens of thousands of context dimensions onto a smaller number (eg 300). We assume that this representation is the only given information about documents and the words therein.

3.1 Semantic Clusters

Now that we have transformed the documents and terms in a higher order non-Euclidian space we can use this to aid our further analysis of the causal relations of the information. Consider the inner product of one document vector upon another. Clearly, documents dealing with similar topics will have a large degree of common words and thus the inner product will be relatively large. Conversely, dissimilar documents will yield low magnitude inner products. This means that we expect documents vectors concerning the same topic would be nearly parallel, and those dealing with different topics would be nearly orthogonal. The goal of LSA is to identify clusters of nearly parallel vectors and thereby discover underlying topics in the library.

In reality, documents and terms will often be spread over several topics; thus, some topics will not be as sharply defined as others because their representative document vectors may be “less parallel”. Moreover, some topics will have a vast number of documents while other topics have relatively few. However the LSA analysis can take advantage of the redundant nature of the input to minimize the effect of such errors. What we mean by redundancy is that documents falling under the same topic will be represented by vectors having many similar entries. Thus, a few errors are not likely to have a significant impact on the analysis.

The analysis for comparing two documents is achieved by calculating the semantic distance between the two vectors in the ***Xhihat*** matrix that represents the documents (columns). Resulting in a value between –1 and 1, where 1 denotes an identical topic contained in the two documents and –1 that no overlap in concepts exist. The semantic distance normally is calculated as the cosine of the angle between the two vectors. Other metrics used were the Pearson and Spearman ranking functions (Dumais, 1996)**.** Thus the matrix ***Xhihat ′Xhihat*** contains the document-to-document dot products. Here the *i ,j* cell of ***Xhihat ′Xhihat*** is obtained by taking the dot product between the i and j rows of the matrix ***DS***. So one can consider rows of a ***DS*** matrix as coordinates for documents, and take dot products in this space.

Spatial Noise reduction techniques and cluster identification

Spatial noise is of key importance when using LSA as defined in (Dumais, 1996). We wish to create an initial term document matrix that more truly reflects the topics contained within the hypertext documents. Our experiments are based on three variations of the input matrix. The first matrix, ***M*** was a set of words that occurred most often in the document corpus. The second ***D*** was a subset of ***M*** that could be found in the Oxford English dictionary (as the query involved terms from the English language). The final matrix ***N*** was the matrix of non-dictionary terms found in the document corpus D.

Now that we have square matrixes ***M,N,D*** we need analyze them to identify possible clusters. This is achieved using the following iterative algorithm comprising of two main stages.

Let n= dimension of square input matrix

Let C be set of possible clusters $C = \{C^1 ... C^m\}\ x, y = [0..n]$

Let δ represent the minimal computed LSA similarity value between strong members within a cluster, and let γ be the LSA value that denotes the threshold where a document is not a member of a cluster.

Stage 1) **Identify Seed Clusters**

$$\textit{If}\ lsa(i,y) > \delta \quad lsa(i,y) \in C_i \tag{1}$$

The initial stage of the algorithm iterates over the square input matrix in order transcending each column and seeing if the LSA value of the document is greater than the upper LSA value δ and for each occurrence of this condition (1) places the document in a cluster C_i.

Stage 2) **Merge Clusters incrementally**

If $x \in C_i$ AND $y \in C_j$ AND: $\min(lsa(x,y)) > \gamma)$ (2)

Unite Clusters (C_i, C_j) i = j = 1 : $lsa(C_i, C_j) = 1$

MergeConstraints: $\delta < lsa(x,y) > \gamma$ (3)

$Max(lsa(x,y) < \gamma)$ (4)

The second stage of the algorithm looks for transitivity between initial clusters. That is to say if we have found two possible clusters C_1 and C_2 there is a chance that they are really one cluster. To check for this possibility we look at the LSA values between C_1 cluster members and the first member of C_2 (2), if this values satisfies (3) which takes account of ambiguity or weak LSA scores (Deerwester *et al*, 1990) then we merge C_1 and C_2

3.2 ANALYSIS OF AMBIGOUS QUERYS

In this section we look at two forms of ambiguity found in natural language; Synonymy and Polysemy. For illustrative purposes we will firstly look at two simple examples. Following this, we will present some real world examples using the same approach. To begin with we generate a document-term matrix by using a given query to identify a collection of related documents using from the Google search engine. We take the top ten results for each query via the Google web-service API. We then retrieve the corresponding documents from the web and pass these documents to our LSA clustering process.

The query “cook” is a polysemous term, which might be submitted by a user when searching for different categories of information. Obvious categories for this term include “cooking food”, the “Cook islands” and “Thomas Cook Holidays”. Each category represents a possible interpretation of the given query term, and we wish our clustering algorithm to identify similar clusters.

Table 1. “Preparation of food” Cluster

Address
http://www.foodsubs.com/
http://www.cooksillustrated.com/
http://www.bawarchi.com/
http://www.canb.auug.org.au/~millerp/cook/cook.html
http://www.everydaycook.com/

Table 2. “South sea islands” Cluster

Address
http://www.cook-islands.com/
http://www.cinews.co.ck/
http://www.everydaycook.com/

Table 3. “holiday information” cluster

Address
http://www.thomascook.com/
http://www.thomascook.co.uk/
http://www.captcookcrus.com.au/

Table 4. unclear cluster

Address
http://www.co.cook.il.us/
http://www.cookreport.com/index.shtml
http://www.inquisitivecook.com/
http://www.montecook.com/
http://www.cookgroup.com/
http://www.cooksgarden.com/
http://www.music.indiana.edu/muslib.html

The query “cook” was used on our clustering process and the following clusters were identified. The first cluster (see table 1) represents the “preparation of food” interpretation, table 2 represents the “south sea island” interpretation and table 3 represents “holiday information”. The results in table 4 have no significant similarity between them and effectively represent mutually independent pages of information. For example, "*MONTEcook.com*" is a personal web site and is unrelated to the other listed sites.

All pages in each of the three identified collections listed in tables 1-3 have a minimum LSA similarity score of 0.9. Furthermore, the maximum similarity between any two pages of different clusters is less than 0.9. In table 5 below, we present average similarity scores between all pages in each of these three collections. (We don't present results for the 4th table as these pages are effectively unrelated to any others.

Table 5. Table of correlation between clusters

	Preparation of food	Cook Island	Holiday
Preparation of food	0.996		
Cook Island	0.72	0.95	
Holiday	0.68	0.87	0.93.

As can be seen from the Table 5 above, the elements within the three clusters have very high Pearson values (0.996,0.95,0.93) indicating a high degree of similarity within each cluster. Interestingly the *cook island* cluster and *holiday* cluster have a stronger correlation of .72 than the weaker 0.68 relationships between "food preparation" and "holiday" clusters.

4 EXPERIMENTAL RESULTS

After seeing how our techniques can be used to remove ambiguity with presentation information that contains multiple target means, we now turn to more real world analysis. The objective of these experiments is to examine the merits of using our LSA clustering techniques in resolving and filtering noise from result sets. By identifying semantically similar groupings and removing noise we filter out ambiguity and deliberately misleading references. The initial sample set of queries is drawn from the top 10 user query's (Google Inc, 2003). *Britney spears, harry potter, matrix, shakira, david beckham, 50 cent, iraq lord of the rings, kobe Bryant, tour de france.*

Our initial set of experiments is centered on the accuracy of our clustering analysis as we vary the input set. (Deerwester *et al*, 1990) from the formation of LSA theory identified the limitations of large matrices in extracting accurate similarities. We ran three experiments varying our initial sample set. We set the singular vector reduction constant r = 5, with the values of γ=-0.4 and δ=0.9. Experiment 1 comprises of all valid lexicons retrieved from the resulting web documents. The second experiment limited the initial data set to only words taken from an English dictionary. The third experiment contained only non-dictionary words. Experiments 2,3 as they used only a subset of the sample data were subject to anomalies such as when no dictionary or non-dictionary words could be extracted. Our experiments account for these anomalies by assigning their computed similarity LSA value to 0. Our clustering algorithm as defined in section 3.1 has three classifications, first are cluster members or references that belong to a topic, second are ambiguous pages that we can't classify using our techniques. The final classification is pages that are not related to any topic in the input set, i.e pages with a LSA value between –1 and γ.

Running our algorithm against the top 10 results from Google user query's for 2003 resulted in accurately determine 77% of the pages into semantically similar cluster, 21% of the data did not form clusters. Only 2% of the data were determined to be anomalies.

In an attempt to create a reference point to compare the computer calculation against, we solicited a five regular human Internet surfers to identify semantically similar groupings from the raw data. This process is similar in fashion as previous (Landauer *et al*, 1998) LSA validation work. The human participants were also asked to identify spurious topics not related to a query i.e. one of the results from the "Britney spears" query points to http://www.google.com/jobs/britney.html, which refers to the Google spelling corrector and has no related meaning to the topic. The surveyed people were also asked to classify the data themselves into semantically similar groups. Comparing this information validation work with the results from our clustering algorithm we found that 62% of the time humans had similar grouping as the algorithm. However humans only determined 5% of the data as non-deterministic and 3.4% of the data as anomalies.

The experiments were rerun for the top 100 results for each of the query's. 37% where determined to be non deterministic with 62% grouped into clusters and only 1% classified as anomalies. As can be seen the non-deterministic classification has grown, this was expected and is in line with other LSA work. We currently don't have corresponding human verification on these results.

The most interesting findings from our initial work centers on experiments 2 and 3. By selectively reducing the input set over the same number of documents by dividing them into dictionary and non-dictionary terms we can increase our accuracy. Experiment 2 "dictionary words" had similar findings to experiment 1 "full data set" with no clear difference being evident. For experiment 3 with the results for the 10 query 10 results configuration of the experiment, we reduced the nondeterministic percentage to 11% while maintaining the anomalies at 2%. When compared with human evaluation we see a clearer division between items that are semantically similar and those that are not. This trend continued as we increased the data set to contain 100 results for each query. Our non-deterministic values reduced themselves to 17%, however the anomalies also increased to 16%.

5 CONCLUSION

In this paper we presented a new clustering algorithm for finding semantically similar web pages. Our techniques borrow from early LSA work, and the pioneering work of Deerwester to resolve natural language ambiguity. We have not examined the computation cost of using this approach for wide scale Internet usage. A detailed description our algorithm providing a brief overview to single vector decomposition that is further backed up by a set of empirical web based examples. The results of applying our techniques to an initial data set using popular search queries from the Google search engine are presented. Our initial findings show promise of using LSA techniques for cluster identification; however as expected LSA introduces a number of new problems when dealing with large data sets. This is evidenced through a loss in accuracy in cluster identification. Our findings have shown that by reducing the input data to the LSA process and therefore into our clustering algorithm we can alleviate the effects of increased data sets. We have shown that one effective approach is to reduce the input set to use only non-dictionary terms. This has been shown to result in cluster determination accuracies similar to those obtained using the smaller data sets.

REFERENCES

Tim Berners-Lee et al. 2001. The Semantic Web, *Scientific American, May 2001.*

Deborah L. et al. 2003. Proposed RecommendationW3C *OWL Web Ontology Language Overview.*

T. R. Gruber 1993. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220.

Deerwester et al 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 1990, 41(6), 391-407

G.W Flake, et al 2002. Self-Organization and Identification of Web Communities. *IEEE Computer.*, 35(3) 66-71.

Klienberg 1998. Authoritative Sources in a Hyperlinked Environment. *In Proceeding of the ACM-AIAM Symposium on Discrete Algorithms.* Pp. 668-667

Peter Pirolli et al. 1996. Silk From a sow's ear: Extracting usable structures from the web. *In Proc. ACM Conf. Human Factors in Computing Systems,* CHI. ACM Press.

Ravi Kumar et al. 1999. Tawling web for emerging cyber-communities, *Proc 8th International World Wide Web Conf.*

Page, L. Brin, S 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *7th International World Wide Web Conference.* Brisbane, Australia, 14-18

Yossi Azar et al. 2001. Spectral analysis of data. *Symposium on Theory of Computing* pp619-626

Dumais, S. T. 1996. Using LSI for information filtering: TREC-3 experiments. In: D. Harman (Ed.), *The Third Text REtrieval Conference (TREC3) National Institute of Standards and Technology Special Publication* , 1996.

H. Papadimitriou, et al. 1998. Latent Semantic Indexing: A probablistic analysis. *In Proceedings of Symposium on Principles of Database Systems (PODS)*. ACM Press

Dimitris Achlioptas et al.2001. Web Search via Hub Synthesis. *IEEE Symposium on Foundations of Computer Science,* pp500-509,2001.

Landauer, T. K. et al 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Google Inc. 2003. Google Zeitgeist; Search patterns, trends, and surprises.