RESEARCH ARTICLE OPEN ACCESS

Document Retrieval System, a Case Study

Mahmood Alfathe,Safa Al-Taie
Electrical and Computer Engineering DepartmentFlorida Institute of Technology, Melbourne, FL 32901, USA

ABSTRACT
In this work we have proposed a method for automatic indexing and retrieval. This method will provide as a result the most likelihood document which is related to the input query. The technique used in this project is known as singular-value decomposition, in this method a large term by document matrix is analyzed and decomposed into 100 factors. Documents are represented by 100 item vector of factor weights. On the other hand queries are represented as pseudo-document vectors, which are formed from weighed combinations of terms.

***Keywords*–**Indexing, Retrieval, Speech, Text

I. INTRODUCTION

This approach tries to overcome the problems of term matching retrieval by statistically treating the unhealable term-document association data. In this proposed method it is assumed that there is some underlying latent semantic structure of data that is may be have been obscured by the random distribution of words. This latent structure has been estimated by using statistical techniques to eliminate the “noise”.

The Latent Semantic Indexing (LSI) which we have used in this project is the singular-value decomposition. We have constructed a “semantic space” by taking a massive matrix of term document association data, where documents and terms that are closely associated have been placed together. With the singular-value decomposition we could arrange the space as all associative patterns of the data and neglect and ignore all he miner, less important effects. At the end most of the terms that have not appear in the document still end close to the document. The position in the space appeared as a kind of indexing. At the end the position of the term in the space of the document determines if the is close enough to the document or not.

Insufficiency Of Current Automatic Indexing And Retrieval Methods

A fundamental deficiency of current information retrieval methods is that the words searchers use often are not the same as those by which the information they seek has been indexed. There are actually two sides to the issue; we will call them broadly synonymy and polysemy. We use synonymy in a very general sense to describe the fact that there are many ways to refer to the same object. Users in different contexts, or with different needs, knowledge, or linguistic habits will describe the same information using different terms. Indeed, we have found that the degree of variability in descriptive term usage is much greater than is commonly suspected. For example, two people choose the same main key word for a single well-known object less than 20% of the time [5]. Comparably poor agreement has been reported in studies of inter-indexer consistency [6] and in the generation of search terms by either expert intermediaries [7] or less experienced searchers [8] [9] . The prevalence of synonyms tends to decrease the "recall" performance of retrieval systems. By polysemy we refer to the general fact that most words have more than one distinct meaning (homography). In different contexts or when used by different people the same term (e.g. "chip") takes on varying referential significance. Thus the use of a term in a search query does not necessarily mean that a document containing or labeled by the same term is of interest. Polysemy isone factor underlying poor "precision". The failure of current automatic indexing to overcome these problems can be largely traced to three factors. The first factor is that the way index terms are identified is incomplete. The terms used to describe or index a document typically contain only a fraction of the terms that users as agroup will try to look it up under. This is partly because the documents themselves do not contain all the terms users will apply, and sometimes because term selection procedures intentionally omit many of the terms in a document. Attempts to deal with the synonymy problem have relied on intellectual or automatic term expansion, or the construction of a thesaurus. These are presumably advantageous for conscientious and knowledgeable searchers who can use such tools to suggest additional search terms. The drawback for fully automatic methods is that some added terms may have different meaning from that intended (the polysemy effect) leading to rapid degradation of precision [1]. It is worth noting in passing that experiments with small interactive data bases have shown monotonic improvements in recall rate without overall loss of precision as more indexing terms, either taken from the documents or from large samples of actual users' words are added [2] [3] . Whether this "unlimited aliasing" method, which we have described elsewhere, will be effective in very large data bases remains to be determined. Not only is there a potential issue of ambiguity and lack of precision, but the problem of identifying index terms that are not in the text of documentsgrows cumbersome. This was one of the motives for the approach to be described here. The second factor is the lack of an adequate automatic method for dealing with polysemy. One common approach is the use of controlled vocabularies and human intermediaries to act as translators. Not only is this solution extremely expensive, but it is not necessarily effective. Another approach is to allow Boolean intersection or coordination with other terms to disambiguate meaning. Success is severely hampered by users' inability to think of appropriate limiting terms if they do exist, and by the fact that such terms may not occur in the documents or may not have been included in the indexing. The third factor is somewhat more technical, having to do with the way in which current automatic indexing and and retrieval systems actually work. In such systems each word type is treated as independent of any other [4]. Thus matching (or not) both of two terms that almost always occur together is counted as heavily as matching two that are rarely found in the same document. Thus the scoring of success, in either straight Boolean or

Coordination level searches, fails to take redundancy into account, and as a result may distort results to an unknown degree. This problem exacerbates a user's difficulty in using compound-term queries effectively to expand or limit a search.

II. THE CHOICE OF METHOD FOR UNCOVERING LATENT SEMANTIC STRUCTURE

The goal is to find and fit a useful model of the relationships between terms and documents. We want to use the matrix of observed occurrences of terms applied to documents to estimate parameters of that underlying model. With the resulting model we can then estimate what the observed occurrences really should have been. In this way, for example, we might predict that a given term should be associated with a document, even though, because of variability in word use, no such association was observed. The first question is what sort of model to choose. A notion of semantic similarity, between documents and between terms, seems central to modeling the patterns of term usage across documents. This led us to restrict consideration to proximity models, i.e., models that try to put similar items near each other in some space or structure. Such models include: hierarchical, partition and overlapping clustering; ultra-metric and additive trees; and factor-analytic and multidimensional distance models,as shown in the survey in [10]. Aiding information retrieval by discovering latent proximity structure has at least two lines of precedence in the literature. Hierarchical classification analyses are frequently used for term and document clustering [11]. Latent class analysis and factor analysis have also been explored before for automatic document indexing and retrieval.

III. CONCLUSION

Assuming there is a set of 'correct answers' to the query. The docs in this set are called relevant to the query. The set of documents returned by the system are called retrieved documents. Precision, is what percentage of the retrieved documents are relevant. Recall, is what percentage of all relevant documents are retrieved. A noisy input is a common problem especially with the OCR techniques where many problems can occur like spelling errors, this suggested method will spell the letters or words correctly.

REFERENCES

[1]. Tougas, Jane E. and Raymond J. Spiteri. "Updating The Partial Singular Value Decomposition In Latent Semantic Indexing". Computational Statistics & Data Analysis 52.1 (2007): 174-183. Web.

[2]. Aswani Kumar, Ch., M. Radvansky, and J. Annapurna. "Analysis Of A Vector Space Model, Latent Semantic Indexing And Formal Concept Analysis For Information Retrieval". Cybernetics and Information Technologies 12.1 (2012): n. pag. Web.

[3]. Atreya, Avinash and Charles Elkan. "Latent Semantic Indexing (LSI) Fails For TREC Collections". SIGKDD Explor. Newsl. 12.2 (2011): 5. Web.

[4]. Phadnis, Neelam and Jayant Gadge. "Framework For Document Retrieval Using Latent Semantic Indexing". International Journal of Computer Applications 94.14 (2014): 37-41. Web.

[5]. Papadimitriou, Christos H. et al. "Latent Semantic Indexing: A Probabilistic Analysis".Journal of Computer and System Sciences 61.2 (2000): 217-235. Web.

[6]. Landauer, Thomas K. Handbook Of Latent Semantic Analysis. New York: Routledge, 2011. Print.

[7]. Foltz, P. W. "Using Latent Semantic Indexing For Information

Filtering". SIGOIS Bull. 11.2-3 (1990): 40-47. Web.

[8]. Aswani Kumar, Ch., M. Radvansky, and J. Annapurna. "Analysis Of A Vector Space Model, Latent Semantic Indexing And Formal Concept Analysis For Information Retrieval". Cybernetics and Information Technologies 12.1 (2012): n. pag. Web.

[9]. Papadimitriou, Christos H. et al. "Latent Semantic Indexing: A Probabilistic Analysis".Journal of Computer and System Sciences 61.2 (2000): 217-235. Web.

[10]. "Latent Semantic Indexing Analysis Of K-Means Document Clustering For Changing Index Terms Weighting". The KIPS Transactions:PartB 10B.7 (2003): 735-742. Web.

[11]. Martínez-Torres, M. R. "Content Analysis Of Open Innovation Communities Using Latent Semantic Indexing". Technology Analysis & Strategic Management 27.7 (2015): 859-875. Web.