

Using concept structures for efficient document comparison and location

Andrew N Edmonds, Scientio LLC

Abstract—A method is discussed for comparing and locating similar documents in a computationally efficient manner by making use of inferred concept statistics, rather than word frequencies. This novel technique uses natural language structures to create a short ‘concept signature’ vector, which locates a document in ‘concept space’. Similar documents can be located in large corpora in $O(\log(n))$ time by making use of this space for indexing. Results from trials with reference and real world data sets are presented, along with a comparison of the method’s document similarity characteristics and the cosine metric.

I. INTRODUCTION

THE explosion in the quantity of searchable textual documents in the past 20 years have created well documented problems in classification and indexing. While algorithms for locating documents that contain key words or even key concepts are well understood and efficient, current algorithms that compare and locate similar documents have less than ideal time complexity characteristics.

Document comparison may mean at least two things: (1) finding documents from a large group that are identical, nearly identical or who share common subsections that are identical or nearly identical, and (2) comparing documents by their inferred internal meaning and content so that documents dealing with similar subjects, with similar structure or similar tone may be located.

In this paper evidence will be presented to show that the algorithm described performs well for use (1) above, and its use for (2) is likely to be a fruitful area for continuing research.

Except where indicated, this document will concentrate on techniques for processing English language documents, though the techniques described can be applied to the vast majority of languages, where a WordNet exists.

II. THE PROBLEM OF SIMILARITY

The definitions of the phrases ‘nearly identical’ and ‘similar’ as used above are a key element of the document comparison process. From a user’s perspective these concepts are vague and dependant on the context of use.

Manuscript received November 14, 2006. This work was supported by Scientio Llc

A. N. Edmonds is with Scientio Llc, Haydon house, Woburn Sands, MK17 8RX UK (phone: +44-1908-766151; fax: +44-1908-766193; e-mail: andy@Scientio.com, web <http://www.scientio.com>). A. N. Edmonds is also a visiting fellow at the University of Buckingham, Buckingham UK,

However it can be surmised that most users of document management systems would prefer a system that ranked returned documents by similarity of meaning, rather than word statistics that may be unrelated to the content. The technique presented here makes use of the inferred concepts in a document, and natural structures in those concepts. Though we do not present evidence in this paper that our technique ranks similar documents more like a human than other techniques, it is to be hoped that the reader will appreciate on further reading that the method has inherently the right attributes to perform well in this area.

III. REDUCING DIMENSIONALITY

This paper introduces a new technique that has been developed by the author and used in commercial applications.

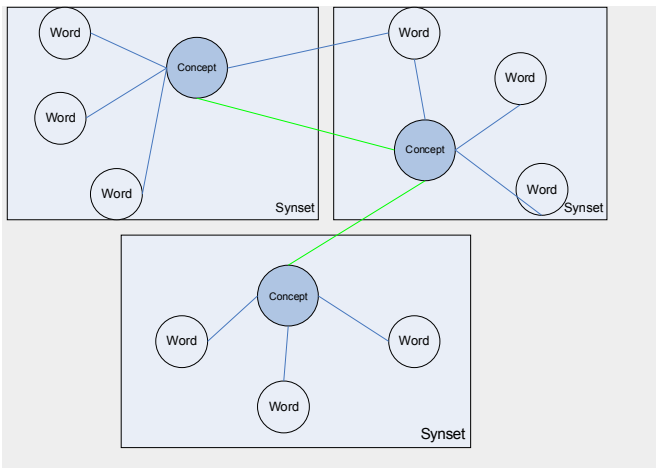
Other techniques for measuring document similarity work, in general, by identifying the words in a document to be analyzed, generating counts of the occurrences, and by performing processing on the counts.

With perhaps 40,000 words in the English language, and a large further number of proper nouns to consider, the dimensionality of the space in which comparisons based on words take place is very large. It is this dimensionality that forces the requirement for very well crafted algorithms. It is also the main barrier to developing lightweight commercial applications in this area.

There are methods for reducing the size of this space; for instance, removing stop words or stemming, but the reduced space is still very large. What is needed is some way to organize words into structures that will permit radical reduction of the number of dimensions.

This paper describes such a method, which makes use of natural structures in language to reduce dimensionality.

WordNet [1] is a freely available database of the English language encompassing the functions of a dictionary and a thesaurus. Like a thesaurus, it organizes words by the concepts to which they relate. These concept definitions, known as ‘synsets’, are the raw material of WordNet. Each synset contains the words that might be used in text where this concept is meant, the definition of the concept, and various links of different kinds to other synsets.



A particular English language word may belong to multiple synsets, corresponding to the different meanings many words can hold in different contexts. In addition to the synsets, WordNet supplies look-up tables that can be used to identify word variants, so that, for instance, 'running' and 'ran' can be associated with the verb 'run'.

The links to other synsets supplied in a synset definition are of vital importance in our analysis. Since they are applied to synsets, they apply to a *concept* rather than a word.

So, for instance, one kind of link connects synsets with opposite meanings. Each word associated with the synsets so linked can, at least under some circumstances, be considered to be the opposite of the words associated with the other synset.

The kinds of links supplied are dependent on the synset and the associated concept. Two common link types are associated with the linguistic ideas of *hyponymy* [3] and *meronymy* [4].

A hyponymy link associates two synsets where one concept is more specific than the other, or looked at the other way round, one concept is more general than the other. Considering noun type concepts, a concept will be a generalization of several concepts, and a specialization of one concept. As a result concepts fall naturally into hierarchies.

Whereas hyponymy encapsulates an '*is a kind of*' relationship, meronymy encapsulates an '*is a part of*' relationship. Thus, for instance, a finger is a part of a hand, which is a part of an arm, which is a part of a body, and so on. Again there are natural hierarchies in these relationships.

Note that both link types are directed.

Using hyponymy and meronymy relationships, and making the assumption that the authors of WordNet have correctly

and completely detailed all the relationships that exist for common English words, it is possible to create a model of the English language.

There are two interesting attributes of this model:

- 1) It is composed of a number of Directed Acyclic Graphs.
- 2) There are only 11 root nodes, corresponding to the eleven elemental concepts identified by WordNet's authors that cannot be further generalized.

To recap then, using the information in WordNet, a massive data structure can be created, containing the words of the English language and the concepts to which they relate, organized into a collection of trees.

Given a common English language word, the associated concepts can be looked up in this structure, and then paths can be traced from each concept to one or more of these 11 root concepts. It is this structure which is exploited to reduce dramatically the number of dimensions in which documents are compared.

IV. USING THE CONCEPT TREES

The key idea of this paper is that these static trees representing words and concept hierarchies might be used in an active fashion. This was prompted by their similarity to Neural Networks, which are also DAGs.

An example algorithm for making use of the trees is as follows:

For each word in a document:

Associate an arbitrary activation value i , with this word.

Find each concept associated with the word (count = c)

Add an activation of i/c to each connected concept node.

Recursively, for each root node

Traverse the trees summing the activation values found at each concept node, creating one sum value.

The above is one of many potential algorithms that might be created to 'activate' the tree structures.

The algorithm creates an 11 dimensional vector representing the document's location in 'concept space'. One dimension of the vector is allocated to each of the

concept trees. Depending on the application the author will often also add other dimensions, summarizing, for instance, the document's size, or the numeric content.

This space, with various extensions, has proved to be a remarkably accurate measure of similarity between documents. It has the following characteristics:

- 1) Documents are represented by small fixed-length numeric vectors. The documents themselves are not needed for further processing.
- 2) The vectors can be treated as a Cartesian space and simple geometric methods can be used to measure similarities
- 3) Well understood algorithms exist that can locate neighbors in such a space in $O(\log_n)$ time.
- 4) Processing of documents is $O(n)$ in document length.
- 5) A variety of data mining and clustering techniques developed for numeric data can be used in concept space.

In the rest of the paper we will look at the uses to which this vector might be put, how it might be used to compare and locate similar documents, and some experimental results detailing the methods effectiveness.

V. SOFTWARE EMBODIMENT

A software class library called *ConceptMap* [7] has been created to embody the above methodology. This was coded in the language C# using the .Net 2.0 common language runtime.

This has three sections:

- 1) An object-oriented model of WordNet's data structures
- 2) A document parsing object that reads a document and creates the location vector.
- 3) A KD Tree [2] based indexing structure used to hold the vectors and locate similar documents in large corpora.

The software is used in two modes that can be combined in any order:

- 1) Document parsing and addition to the index
- 2) Locating similar documents to a given target in the index.

To use the software the user must supply a collection of documents and unique keys for each. When similar documents are required the user can present either a new document or an existing key, and an ordered list of document's keys is returned, along with distance information. New documents can be added to the index at

any time.

VI. TRIALS ON DATASETS

ConceptMap has been evaluated on several large datasets, and is in commercial use as a means of finding near duplicates in a 2.5 million document dataset, consisting of legal and constitutional text in Spanish.

The following trial was performed on a cut down version of the Reuters 21578 dataset. This consists of 8599 news stories, individually labeled and annotated. The stories are short, averaging at 150 words each. Many similarity measures have problems with short documents, so this is an interesting trial. Also, many of the news stories contain financial information written in a very similar format, thus representing a difficult subject for any similarity measure. The dataset contains no exact duplicates.

The first part of the trial was to ensure that this software would retrieve an identical copy of the story reliably. Thereafter corrupted versions of the stories were created by truncating 5% of the text from the start, and in a separate trial from the end. The system was then asked to find the nearest neighbors to these corrupted documents, and the position in the list of the uncorrupted version was noted.

TABLE I
REUTERS DATA SET TRIALS

Rank in search by which document was found	No truncation %	Truncation at front %	Truncation at rear %
1	100	86.03	71.14
2		91.68	79.95
3		93.66	83.52
4		94.89	86.06
5		95.68	87.59
6		96.12	88.84

Thus the algorithm reliably found identical documents, and showed good performance with corrupted documents.

The response from the commercial application of this system has been good. The users have expressed satisfaction with the ability to find similar documents, and report search times of less than 3 seconds to retrieve the nearest 10 documents to a new document, using standard single CPU servers.

VII. COMPARISON WITH OTHER SIMILARITY METRICS

A reference database [5] of cosine similarities between pairs

of documents drawn from the NIST WT10G [6] dataset has been created. Cosine similarity is the cosine of the angle subtended between two documents' word vectors when considering only those words that are found in one or both documents. While the cosine similarities represent a well understood means of comparing two documents, in order to find similar documents in a corpus of n documents, $n(n-1)/2$ similarities must be calculated. Thus such an algorithm has time complexity of $\sim O(n^2)$ and is impractical for large data sets. The creators of the database report that their calculations required 3 weeks of computer time.

Both Cosine Similarity and ConceptMap can be reliably used to identify duplicates in large data sets. When considering similar documents, however, it is clearly of interest to discover how the two measures compare.

ConceptMap and cosine similarity measure very different aspects of a document. As discussed previously, document similarity, when viewed from a user's perspective, is a subjective quantity.

A simple method for measuring the agreement of the two similarity measures was chosen. For a large set of 'pivot' documents drawn from the WT10G dataset and used in []'s calculations, the nearest neighbors, measured using cosine similarity, were collected.

The neighbors were then ordered in decreasing cosine similarity.

All the documents that had been used by in the generation of the reference database were drawn from the WT10G dataset and loaded into ConceptMap. ConceptMap was then required to locate the nearest neighbors in 'concept space' to each of the 'pivot documents', again ordered in descending similarity.

For each 'pivot document' we thus had two lists of nearest neighbors using the different measures of similarity.

It was decided to consider the first 10 of each list, and to measure the similarity between the two sets of neighbors using Jacquard's coefficient. The results were:

TABLE 2
WT10G DATASET TRIALS

Pivot documents	29568
Average Jacquard's coefficient	0.22
Standard deviation of Jacquard's Coefficient	0.21

A Jacquard's coefficient of 0.22 is equivalent to ~ 4 documents in common between both lists.

VIII. DISCUSSION AND FURTHER WORK

The algorithm presented is clearly an effective and efficient

method for identifying similar documents in large corpora.

As would be expected, there is some, but not complete agreement in the ordering of neighbors between our concept-based technique and the benchmark word frequency technique.

Clearly, very interesting research could be performed comparing these two measures from a human usability point of view. No such research has yet been undertaken, yet the following points should be made.

Because a concept lookup is performed as part of the measure generation, documents that have been revised by replacement with synonyms score as similar with ConceptMap, and different with word based techniques. For instance *'The cat sat on the mat'* and *'the feline squatted on the rug'* would score high similarity with ConceptMap, and low cosine similarity. ConceptMap would appear to be more useful for identifying revised versions of existing documents.

ConceptMap could be described as a 'bag of concepts' technique in that it generates an aggregate measure of concepts within a document and the order of presentation of concepts does not affect the location in 'concept space'. The same arguments apply against 'bag of concepts, as against 'bag of words' in that both discard much of the information in a document. The same argument in favor seems also to apply, the fact that they work in practice.

ConceptMap takes advantage of the concept structures inherent in the WordNet model of English. One might argue that the concept structures in WordNet are entirely arbitrary, and so any dimensionality reduction obtained through them is worthless, and the apparent value as a similarity calculation tool is illusory. A full answer to this argument would require more space than is available, but the structures embedded in WordNet are not arbitrary, they represent the mental constructs we use in speech, and their relationships are remarkably robust across cultures and languages. Nobody would deny, I hope, whatever culture they came from or language they used that a pony was a kind of horse or running was a kind of movement. Similarly the top level concepts are those that cannot be generalized any further in any language or culture. Evidence for, though not proof of this assertion can be drawn from the fact that ConceptMap has been implemented using two independently created WordNets for English and Spanish with similar performance in both.

There are at least two interesting avenues for future work.

As discussed, any individual word may be associated with several concepts. When this word is used only one concept is intended, and yet ConceptMap excites all the concepts equally when analyzing a document. There is a lot of work being performed [8] for other purposes to try to create

structures that would enable the context to be identified. Use of this information might improve usability.

The other avenue is the use of ConceptMap to find documents that are similar in topic, structure, argument or tone. These quantities are very nebulous, and yet very attractive. Some work has been undertaken in this area using internet derived weblog text, and it is clear that for these applications not only the root nodes of the concept trees, but the first generation children need to be used to build a suitable concept space. The author is looking for suitably marked up data sets to take this research further.

ACKNOWLEDGMENT

A. N. Edmonds would like to thank Abdur Chowdhury and Eric Jensen of The Illinois Institute of Technology, Information Retrieval Laboratory and Hongbo Du of the University of Buckingham for their help in supplying data and discussing this work.

REFERENCES

- [1] WordNet An Electronic Lexical Database, Edited by Christiane Fellbaum, MIT Press, 1998, ISBN: 0-262-06197-X
- [2] Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (Sep. 1975), pp 509-517.
- [3] Hyponym. (2006, August 27). In Wikipedia, The Free Encyclopedia. Retrieved 20:32, November 12, 2006, from <http://en.wikipedia.org/w/index.php?title=Hyponym&oldid=72204125>
- [4] Meronymy. (2006, October 24). In Wikipedia, The Free Encyclopedia. Retrieved 20:29, November 12, 2006, from <http://en.wikipedia.org/w/index.php?title=Meronymy&oldid=83358669>
- [5] Duplicate Document Detection Test Collection, Abdur Chowdhury & Aleksander Kolcz, (February 5, 2005), Retrieved October 25 2006, from <http://www.ir.iit.edu/~abdur/Research/duplicate-collection.html>
- [6] NIST (2006). Trec home page. Available: trec.nist.gov/. Accessed November 12 2006.
- [7] ConceptMap, Scientio LLC (2005-2006) <http://www.scientio.com/conceptmap.aspx>
- [8] [1] Liu, H. & Singh, P. (2004) ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, Volume 22, pp 211-226 Kluwer Academic Publishers.