EinCite:
The Key to Document Similarity

Background & Motivation

The Web is big. Really big. Mind-bogglingly big. Search engines are small, very small.

A 1999 study determined there were at least 800 million Web pages, an increase of 40% over the previous year (Lawrence and Giles 1998, 1999). How can a user find interesting documents? The answer previously was to use a search engine. The same study revealed the top six text-based search engines only indexed 16% of the pages, a drop of 17% from the previous year. We need something better.

92% of Canadians surveyed (PriceWaterHouseCoopers 2000) listed searching for information as one of their top Internet activities. Most approaches to solve this problem fall into four major areas: index more, address the currency problem, index better, and match queries to results better. The EinCite project focuses on the last two issues: indexing better/better hypertext document classification and matching queries to known hypertext documents better. If you cannot ascertain that one document is similar (or not) to another document in terms of its content, it is difficult to return accurate, meaningful search results on the web. EinCite will build better search engines by initially investigating methods of improving the matching of documents to queries and deciding which documents are similar to other documents utilizing semantic relations in the WordNet lexicographic database.

EinCite (Phase 1): The 30-Second Summary

What if a document had a fingerprint? What if you could tell how similar two fingerprints were? What if you could group fingerprints by their similarities? Welcome to EinCite.

EinCite is a multi-phase document classification, information retrieval, and intelligent agent system for retrieving interesting hypertext documents on the basis of a fingerprint calculated by using semantic relations from WordNet to determine a document's subject. In the document classification phase (Phase 1), each concrete noun in a document can be said to belong to a more general class of nouns. These more general classes are aggregated and provide meta-information that can be used to build the document fingerprint and to classify the document.

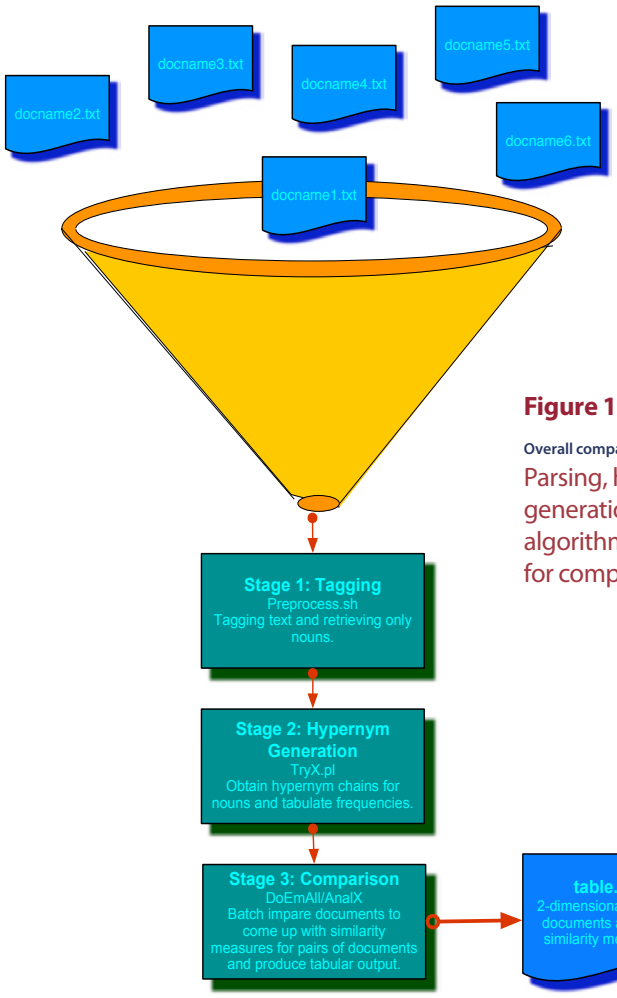


Figure 1:

Overall comparison process

Parsing, hypernym generation, and algorithm application for comparison.

Research Questions

What do we hope to discover?

There are four primary research questions:

1) Can the hypernym semantic relation be used to build a document fingerprint of the key concepts in a document?

2) Can a set of document fingerprints be compared to determine how similar a set of documents are to one another and therefore improve search engine retrieval?

3) Does such a similarity rating compare favourably with expert ratings or average web user's ratings?

4) Can document fingerprinting be applied to the automatic classification of documents as part of the effort to create the semantic web?

IDEAs Lab, Informatics, University of Sussex
More info at http://einiverse.eingang.org/
Michelle A. Hoyle (eingang@sussex.ac.uk)

Document Fingerprinting & Document Comparison

What is a document fingerprint? How are they built? What is a hypernym? How do we compare documents?

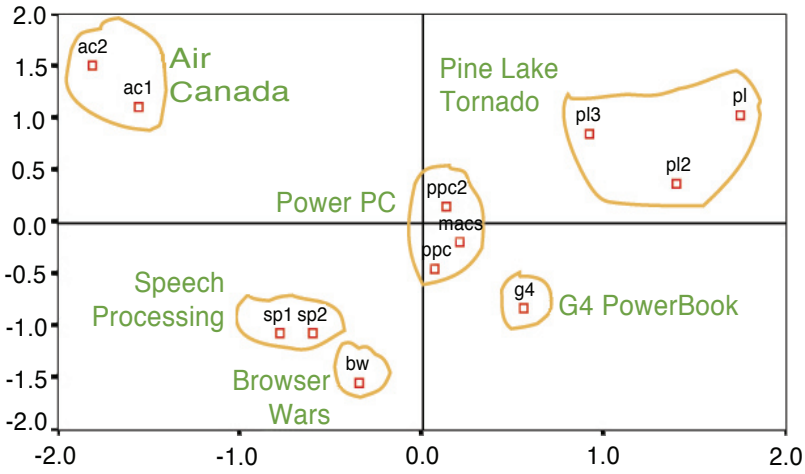
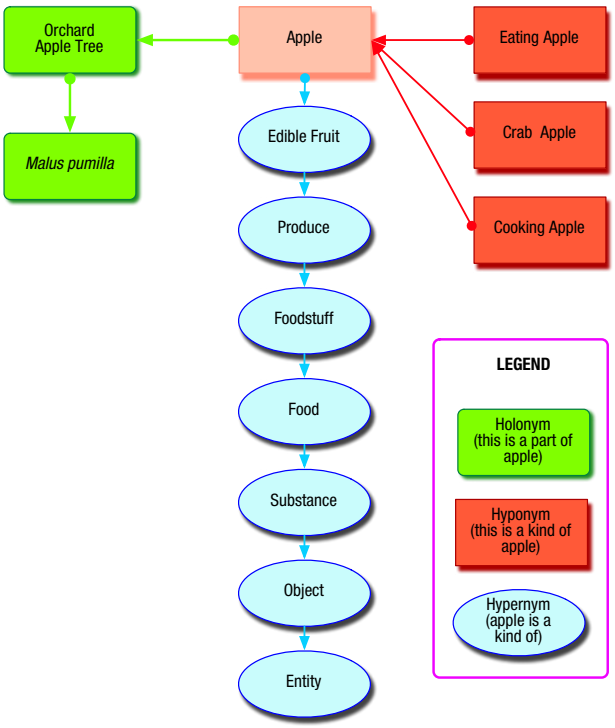


Figure 3: Sample Similarity Clustering
The clusters are circled, with nearest green text indicating document subject. This was using one of the in-house similarity metrics.

Hypernyms are the key to building a document fingerprint, which is a collection of graphs depicting the number and strength of relationships between nouns in the document (See Figure 4). A hypernym of a noun is simply a more general, abstracted idea of the noun, generally denoted as being the "kind of" relationship. Apple is a kind of edible fruit, so edible fruit is a hypernym of apple (See Figure 2). Most search engines are strongly keyword-based, but what if a document doesn't mention fruit at all, but you still want to look for articles about fruit or aren't interested in whether a specific word occurs in a document? The theory is that the use of hypernyms, which abstract the concepts in a document, can be used to build a more general understanding of a document. Will each "fingerprint" be unique as human fingerprints are? Probably not, but that is an open question.

The overall process of generating a document fingerprint is partially depicted in Figure 1 outlining how to compare documents. A document is first digested by a set of natural language parsing tools to tag and retrieve only the nouns by context as hypernym relations are not available for other parts of speech. A hypernym is retrieved, if available, from the WordNet lexicographic database for each noun in the list. Each hypernym could also have its own hypernym which is also retrieved back to a variable depth. So, for example, for "apple," we might retrieve all the hypernyms in Figure 2 up to and including "Food." As nouns are likely to share hypernyms, we increment a count for each shared hypernym. At this point we have enough information to generate a fingerprint, but we continue on with the document comparison process. In Stage 3, the hypernyms, terms, and their associated weights are applied against some formulae, matching up weighted terms and hypernyms to a comparison document. This comparison is bidirectional as some of the similarity metrics are not symmetric. This gives us a similarity matrix of all processed documents.

Figure 2: Hypernyms of 'Apple'
The middle column shows, in descending order, more and more abstract hypernyms or "kind of" relations for "apple." An apple is a "kind of" edible fruit, which is a "kind of' produce and so on.



The Experiments & Results So Far

Prototypes and people. Clustering is grouping similar documents together well, at least with some algorithms.

A set of two experiments have been run using sets of twelve documents, eight people and two automated metrics. The people and the prototypes follow the same sequence of steps: read through each document in pairs and decide how similar each pair of documents is on a scale. The result is a two-dimensional similarity matrix which can then be analysed using multidimensional scaling (MDS) (Borg, 1997) to produce a clustering of how similar

the documents are (See Figure 3). The primary prototype algorithm encapsulates the notion that frequently occurring hypernyms are more representative of the overall document subject and should be more important. The second algorithm uses a simple pairwise comparison with cosine normalization, common in information retrieval tasks. MDS allows us to aggregate the rankings by people to produce an overall clustering, but it does not allow us to easily compare two sets of rankings to see how close they are because placement in the vector space is dependent on the data used to make the space. Adjusting the prototype algorithms is therefore difficult as it is not possible to quantify the level of improvement. Nevertheless, some initial output from the prototypes (e.g., Figure 3) looks very promising.

Future Work

There's much left to explore with generating fingerprints and adjusting formulæ.

Future work lies along four major axes: data analysis; background research; formula modification and implementation; and additional experiments. Given a set of clustering graphs from the MDS process (See Figure 3), other than by visual inspection, how can the graphs be compared to determine how close the automatic classification is to the participants or how the different machine algorithms compare against one another for accuracy? Formula work includes trying other traditional metrics like TFIDF (Salton and Buckley 1988; Spärck-Jones 1972), a variant of a standard information retrieval metric used in many web search engines. In recent years, some similar approaches using WordNet have appeared, including Brezeale's (Brezeale 1999), so more recent literature should be examined. Modifications should also be made to the overall procedure of selecting the hypernyms, a more dynamic approach that can automatically identify when a concept is too vague to be useful, perhaps based on how many terms share the same hypernym in WordNet. Finally, completely generating the fingerprint and using more resultant graph features for comparison.

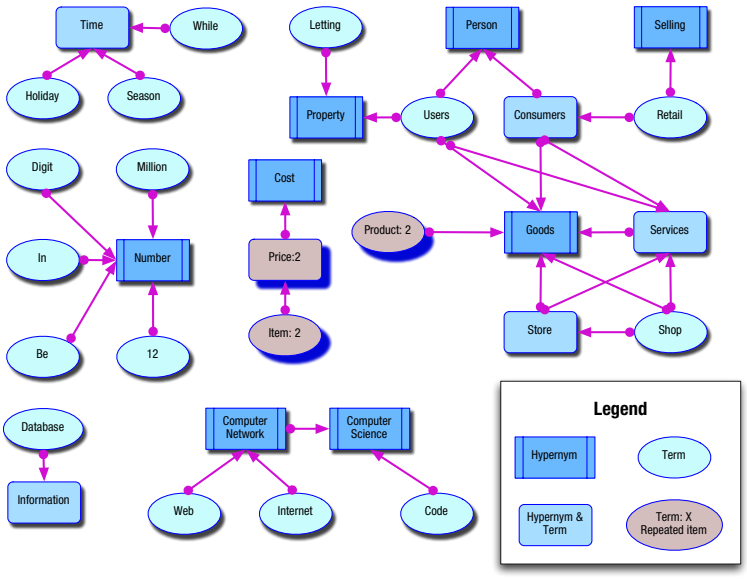


Figure 4: Sample document fingerprint
This fingerprint was built by hand from a short ***Wired*** article. The clusters are hypernym concepts (rectangles) generated by the article's nouns (terms as circles). The larger the hypernym cluster, the stronger the importance of those linked concepts. The largest cluster here is related to sales and goods. This article was about selling.

Selected Bibliography

Borg, Ingwer, and Patrick Groenen. 1997. "*Modern Multidimensional Scaling: Theory and Application.*" Springer Series in Statistics. New York: Springer Verlag.

Brezeale, Darin. 1999. "The Organization of Internet Web pages Using WordNet and Self-Organizing Maps." M.Sc., Department of Computer Science, University of Texas at Arlington, Arlington.

Lawrence, Steve, and C. Lee Giles. 1998. "Searching the World Wide Web." *Science*, April 3, 1998, 98-100.

Lawrence, Steve, and C. Lee Giles. 1999. "Accessibility and Distribution of Information on the Web." *Nature*, 1999, 107-109.

PriceWaterhouse Coopers. 2000. "Canadian Consumer Technology Study 2000." PriceWaterhouse Coopers.

Salton, Gerard, and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24 (5):513-523.

Spärck-Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation* 28:11-21.