Traditional Resources Help Interpret Texts

J. Gelernter
Carnegie-Mellon University
Pittsburgh, PA
gelern@verizon.net

M. E. Lesk
Rutgers University
New Brunswick, NJ
lesk@acm.org

ABSTRACT

Simple word matching between the user query and document is common, as are mis-matches of meaning that occur as a consequence, and errors in recall. These defects in the "bag of words" model are well known, and raising the semantic level of representation will improve retrieval. This can be done by expanding words and user queries using traditional reference sources such as gazetteers and synonym lists or ontologies.

Categories and Subject Descriptors

H.3.1 [**Information Retrieval**]: Content analysis and indexing – *thesauri, gazetteers, subject analysis*

General Terms

Performance.

Keywords

Ontologies, knowledge representation, information retrieval.

1. INTRODUCTION

The defects of purely word-based retrieval have been known for a generation. It focuses our searching on named entities rather than concepts, it makes it difficult to search for particular relationships between concepts, and it suffers from both the ambiguity and synonymy of words in texts. The ease and speed with which it can be done, however, encourage its use. Almost by default, retrieval systems today operate with a "bag of words" approach[1], avoiding all consideration of syntax or of the relationships between words and concepts.

Proper names are handled best in this context; they are likely to be unique and have no synonyms. Nobody has any doubt what is meant by writing *Costa Rica* or *Sir John Gielgud* and nobody is likely to refer to these entities in other words. But proper names are not that much of English. Picking a 19th century book containing 100,000 words, about half are syntactic function words but only about 7,000 are proper names.

Beyond named entities, we find that words are too often ambiguous; is *orange* a color or a fruit [2]? More important, there are many ways of saying the same thing. Tom Landauer reported that people asked to name common concepts (for example, a text editing command) typically come up with half a dozen different words for it [3].

Potentially, the focus on specific named entities may be changing the way we think about intellectual subjects. The focus may be entirely on little bits of specific knowledge, and on isolated rather than connected ideas. Criticizing the way computer retrieval systems steer our thoughts is quite popular now: see for example Carr's article "Is Google Making Us Stupid?" in the *Atlantic* [4]. Literary scholarship, for example, seems not to have gained as much from online texts as one might have expected; the ability to say instantly how many times Milton used the letter K does not shed much light on his ideas (although there is a paper comparing the number of superlatives in Dickens and Smollett).

What can we do about this? Two answers from the retrieval community have been clustering and collaborative filtering. Clustering tries to group related words, and has been implemented in such search engines as *clusty.com;* it has not attracted widespread use. For example, if the *clusty* search engine is given either *Vancouver* or *Melbourne*, in neither case does it quickly indicate a category of replies for the person rather than the city named for them (George Vancouver, Lord Melbourne). Collaborative filtering [5] replaces word searching with collective experience, and it is indeed a completely different base for conceptual representation. However, it requires that numerous people see each item. In a very large research library, especially for older materials, the problem may be precisely to find items that have not been viewed recently.

Recently the "semantic web" [6] has been proposed as a way to make online content more easily useful. All items in a document would be labeled, so that we would know what numbers represent temperatures or prices, what the context of each word is, and so on. This ambitious scheme, extending well beyond words to numerical data, is perhaps beyond the amount of manual work that anyone is prepared to expend on most texts. As of now, we do not have reliable software to do this kind of tagging automatically.

2. REFERENCE WORKS

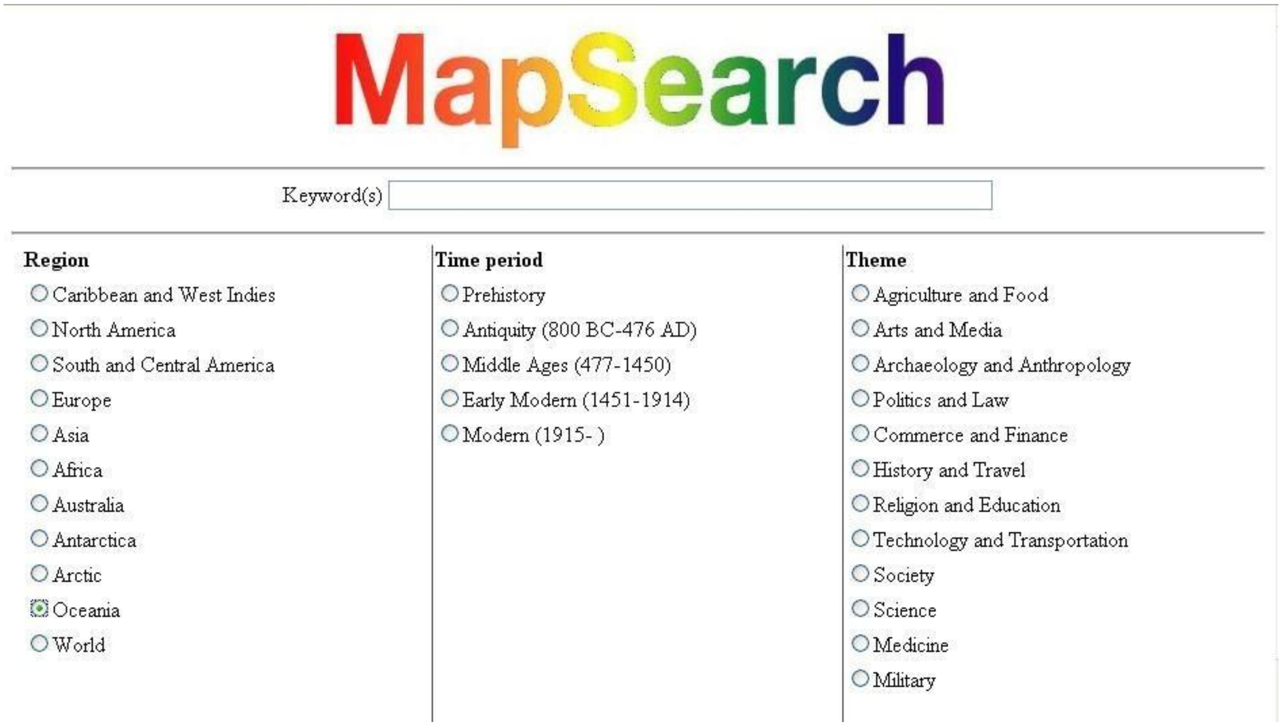
Historically, we have relied on large scale manual efforts to create reference works. These include dictionaries, encyclopedias, thesauri, and classification systems. The best known are probably the *Oxford English Dictionary* [7], [8] and the *Dictionary of National Biography*, but the lack of public attention to library catalogs, thesauri, encyclopedias, gazetteers, and the like [9] does not make them less valuable. These were created to help people use language and library resources, and although not built with computers in mind, they can be used to help programs as well as people. An early example was the use of dictionaries to help with

BooksOnline'08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558- 249-8 /08/10...$5.00.

sense disambiguation [2]. More recently Leidner [13] used gazetteers to recognize geographic entity names.

There is a difficulty in that each of the historical reference works were created separately. Thesauri descend from the work of Peter Mark Roget, and they do not intersect with dictionaries, which descend from Samuel Johnson, Noah Webster, and others. Even now there is no paired dictionary and thesaurus in book form, such that the different senses of a word in the dictionary correspond to the meaning classes in the thesaurus.

WordNet was developed [10] partly to meet this need. However, WordNet is somewhat idiosyncratic, and it seems worth exploring what can be done with more traditional resources. This paper will discuss two cases where we might profitably use a reference work to improve search of books or articles. The first is about maps, the second about text.



MapSearch

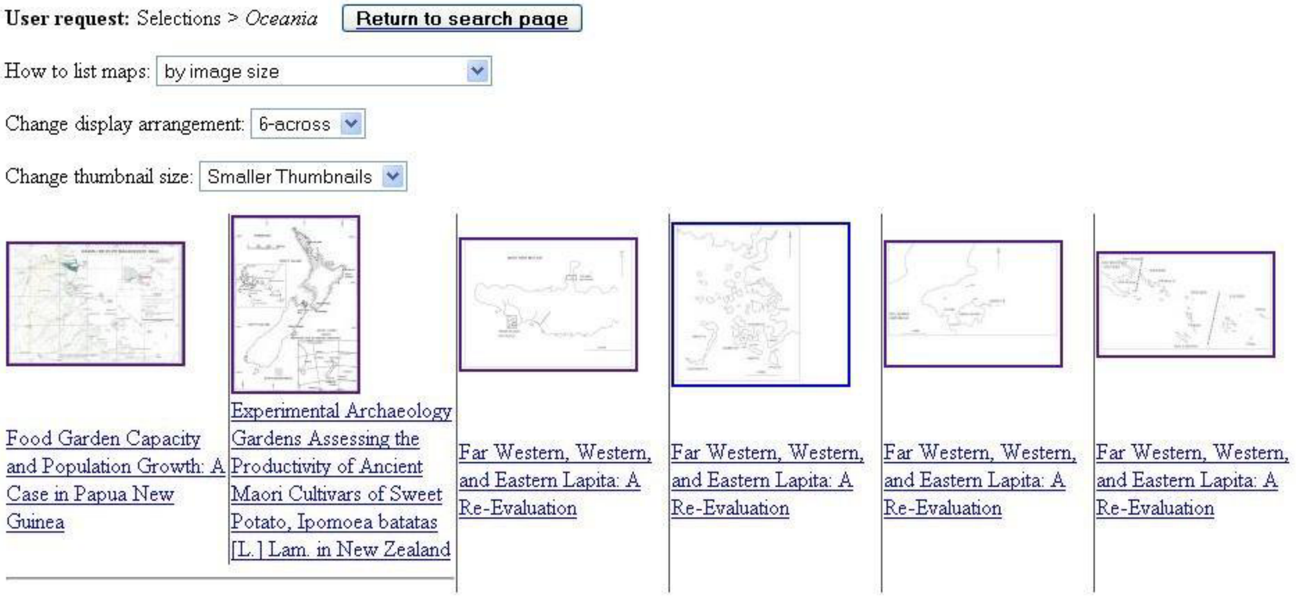
Keyword(s)

Region
- ○ Caribbean and West Indies
- ○ North America
- ○ South and Central America
- ○ Europe
- ○ Asia
- ○ Africa
- ○ Australia
- ○ Antarctica
- ○ Arctic
- ● Oceania
- ○ World

Time period
- ○ Prehistory
- ○ Antiquity (800 BC-476 AD)
- ○ Middle Ages (477-1450)
- ○ Early Modern (1451-1914)
- ○ Modern (1915-)

Theme
- ○ Agriculture and Food
- ○ Arts and Media
- ○ Archaeology and Anthropology
- ○ Politics and Law
- ○ Commerce and Finance
- ○ History and Travel
- ○ Religion and Education
- ○ Technology and Transportation
- ○ Society
- ○ Science
- ○ Medicine
- ○ Military

Figure 1



User request: Selections > *Oceania*

How to list maps: by image size

Change display arrangement: 6-across

Change thumbnail size: Smaller Thumbnails

Food Garden Capacity and Population Growth: A Case in Papua New Guinea

Experimental Archaeology Gardens Assessing the Productivity of Ancient Maori Cultivars of Sweet Potato, Ipomoea batatas [L.] Lam. in New Zealand

Far Western, Western, and Eastern Lapita: A Re-Evaluation

Far Western, Western, and Eastern Lapita: A Re-Evaluation

Far Western, Western, and Eastern Lapita: A Re-Evaluation

Far Western, Western, and Eastern Lapita: A Re-Evaluation

Figure 2

3. CONCEPT SEARCH EXAMPLES

3.1. MapSearch

Lee Giles of Penn State University suggested the problem of extracting maps from journal articles for individual retrieval. In general, such maps are not indexed individually and can be difficult to retrieve. Our research in [11] explored the problem of how to organize these maps. Maps, of course, are characterized by several specific attributes, including the area of the earth covered and the date represented. In addition, most maps used as illustrations in articles are thematic maps, which provide information about some subject, and users would like to be able to select maps based on that topic. Words alone are not always going to be adequate, since, for example, maps about agriculture might refer to crops, soils, names of specific plants, or farming; the word agriculture itself might never appear. So a browsing interface was offered as an alternative, with three columns giving the possible areas, times, and subjects (Figure 1). Figure 2 then shows a result screen of maps retrieved. Note that none of the map or maps texts retrieved by the search on “Oceania” contain the word “Oceania” in their metadata. The reference gazetteer groups New Guinea and New Zealand as shown in the examples and retrieves them in the “Oceania” category.

How was this done? We started by extracting words about the maps: these might come from the caption, the words printed in the map itself, the title of the article containing the map, or the sentence referring to the map. Stop words were removed. These words were matched against some standard reference sources. Both gazetteers and dictionaries were used to index by region to help manage place names that are also English words. For example, although *Toronto* is almost always going to be a city name, and is important enough to be used without any qualification, the word *Bath* is probably an ordinary English word unless there is some indication (at least capitalization) that it is a place name. Although there is a creek named *What* it is particularly unlikely that this word, if found by itself, refers to water.

Maps were classified into date categories mostly numerically, but time words and phrases such as found in the Library of Congress classification schedule for history also were used. One difficulty here was to decide which numbers represent dates, and which dates belong to the map rather than the publication the map is taken from, cites, or is otherwise related to. But again, the use of a reference work enables us to know to what date a period name like Elizabethan might refer.

To perform concept matching by subject, we used classification categories associated with the Library of Congress classification (several classes were combined for convenience), and associated ontology synonyms with each category we made. We weighted certain phrases based on the degree of certainty to which we felt with which they pointed to the particular classification (example: “first aid would be classified in Medicine). Then the per-map work was automatic, so that if a phrase was found in a map caption or associated text, it counted as part of that classification category.

The benefits of using reference sources is thus twofold: to provide an organization of categories that users may become accustomed to (whereas dynamic clustering typically provides a different organization routinely), and to draw upon pre-made lists of words and phrases to improve recall in information retrieval.

3.2 Phrase matching in books

As mentioned before, (a) there are too many words in English, so that sometimes we don't find word overlaps when we need them, (b) the focus of most retrieval systems is on entities and not relationships. Interestingly, chemistry has always had the same problem: retrieval by substance is done well while retrieval by reactions is comparatively poor.

We ran a book through a parser to consider how we might accomplish by phrases. Randomly selected for the example was Francis Parkman’s *History of France in the New World* that is available digitally courtesy of Project Gutenberg, and the Stanford parser. We can then make lists of word pairs, say verb-object pairs. The things that got built (objects of the verb *build*) were *forts, huts, cities, ships, buildings,* and so on. The things that were done to *forts* (verbs with *fort* as object) include *build, occupy, visit, approach, inspect, demolish,* and *abandon.*

We can run the same book through a thesaurus instead of a parser for other sorts of phrases. There is a 1911 Roget Thesaurus available on Project Gutenberg, which is just fine for an 1865 book. If one takes each phrase and replaces the words with the category numbers, we can see that *build fort* is related to *pile breastwork* and *lay bank.* This is not quite as straightforward as it seems, since many words appear in multiple thesaurus categories, and other words do not appear at all. Some degree of sense disambiguation is needed to decide which categories are relevant to this particular sentence, and we will also need further work to decide which phrase combinations are most likely to be useful.

An alternative method of assigning words to categories made use of a collection of library catalog records. Given a few million book titles labeled with Dewey decimal classification, each word can be associated with the Dewey number of the titles in which it appears. Again, this has the problem that common words appear in a great many different categories (*building* appears in more than 600 of the 999 possible categories). But again, by focusing on the most frequent meanings of the words, and looking at phrases which are syntactically parallel and use semantic related words, we can find apparently similar phrases, for example, relating *build fort* and *build city*.

One advantage of the Dewey data is that it is multilingual. For example, the two highest frequency words in titles in Dewey category 910 in English books are *geography* and *travel*. In French they are *voyage* and *geographie*. In Spanish they are *geografia* and *viajes*. In general, however, it appears that the thesaurus does a better job of relating words to semantic ideas. However, since the coverage of the two databases is different, they can be combined to include a greater number of words. The thesaurus has some 50,000 words; the Dewey titles include more than 400,000 different strings, and although some are irrelevant, most are actual words or names.

In principle, one could select book titles by date, and get semantic information that was keyed to the exact date that the book was published. Words like *train* and *post*, for example, have changed their meaning considerably over the centuries. As an extreme example, consider the lines .... *and they, in night Of their ambition, not perceive the train, Till in the engine they are caught and slain.* (Ben Jonson, Sejanus, Act II, 267-269). This probably produces a particular image in your mind, including something about a railway, but it was written by Ben Jonson and involves older meanings of both *train* (scheme, trickery) and *engine* (catch, mechanism). Unfortunately, it is not yet clear that we know how to use resources well enough to do this kind of time-sensitive disambiguation.

Again, though, the key idea is that by using reference works as search intermediaries, we get reproducible groupings of words, and we get groupings of words that people have seen before, that expand the query or the document text and improve retrieval.

3 CONCLUSION

Even in the last century, standard and biographical dictionaries, gazetteers and thesauri that enrich reading have been compiled. Using such sources to mediate between query and document terms will improve concept matching and the recall aspect of information retrieval.

We should be able to balance syntactic and semantic specificity. We can either request exact matches of words and minimal syntactic relation (as we do now) or move towards broader semantic categories and then be able to ask for more syntactic relation while still retrieving a reasonable number of items. If searching of that kind enabled a more conceptual view of retrieval, it would be more of what you would get from the content of a book, and less what you would get from the index. In other words, it would be what the original author wanted us to get from a book, rather than the snippets we are prone to see now.

Long ago, Dick Hamming began one of his books with “The purpose of computing is insight, not numbers” [12] at a time when we didn’t publicly know that the first computing had been about cryptography and thus text. So perhaps we could say today that the purpose of computing is insight, rather than word frequencies.

4 REFERENCES

[1] Schutze, H. 1992. Dimensions of Meaning. Proc. IEEE/ACM Supercomputing, 787-796.

[2] Lesk, M. 1986. How to tell a pine cone from an ice cream cone. Proc. SIGDOC conference, 24-26.

[3] Landauer, T. K., K. Galotti, and S. Hartwell. 1983. Natural command names and initial learning: a study of text editing terms. Comm. ACM, vol. 26, 495-503.

[4] Carr, N. 2008. Is Google making us stupid? Atlantic, July-August. Retrieved August 5 from http://www.theatlantic.com/doc/200807/google

[5] Hill, W., L. Stead, M. Rosenstein, and G. Furnas. 1995. Recommending and evaluating choices in a virtual communityof use. Proc. SIGCHI, 194-201.

[6] Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The Semantic Web. Scientific American, 16 pp. Retrieved August 6 from http://www-personal.si.umich.edu/~rfrost/courses/SI110/readings/In_Out_and_Beyond/Semantic_Web.pdf.

[7] Winchester, S. 1998. The Professor and the Madman. HarperCollins.

[8] Murray, E. 2001. Caught in the Web of Words. Yale University Press.

[9] McArthur, T. 1986. Worlds of Reference. Cambridge University Press.

[10] Fellbaum, C. 1998. Wordnet: An Electronic Lexical Database. MIT Press.

[11] Gelernter, J. 2008. MapSearch: a protocol and prototype application to find maps. PhD thesis, Rutgers University.

[12] Hamming, R. 1962. Numerical methods for scientists and engineers. McGraw-Hill.

[13] Leidner, J. Topnym Resolution in Text. Edinburgh PhD thesis, 2007.