The Technology

Colvin & Company

Greg Colvin
PO Box 1418
Buena Vista, Colorado
81211
•
719-221-2575
greg@colvin.org
www.colvinco.net

Thursday, February 21, 2008

All and Only the Relevant Information

What we have is an information-matching technology that we are initially applying to text retrieval and ad placement. Information-matching goes beyond merely matching words to matching concepts.

This technology is based on representing conceptual content as vectors, called information spectra, and retrieving information based on the similarity of the spectra. This amounts to treating spectra as points in a high-dimensional information space. When an information spectrum represents a query then we can retrieve information by finding nearby points in the space.

Information spaces are created a using psychometric technique known as Classification Space Analysis or Judgment Space Analysis which goes back to Peter Ossorio's early (1964,1966) work. What he did was to work with subject matter experts to create a set of subject-matter classifications, and to identify a set of phrases to characterize each subject matter. Then he had the experts rate the relevance of each phrase to each subject matter. These judgments he gently massaged into a vector space to support information retrieval. The operation of this space he called Judgment Simulation.

The result, replicated and extended over the years by Joe Jeffrey (1976, 1991, 1993), Ossorio (1989), Michael Kurz (1989, 1992, 2000) and others, is a retrieval system that can simultaneously approach 100% precision (finding only relevant information) and 100% recall (finding all relevant information.) In the sixties results this good were simply unprecedented – for all other known techniques the tradeoff was *precision + recall <= 100%.* Newer approaches like Latent Semantic Indexing (Deerwester et al. 1990) are only recently (Hoffman 1999) getting past that tradeoff.

We explain these good results by the fact that a subject matter is the world of a community and communities tend to develop distinctive language to talk about their worlds. This language is only poorly captured by purely statistical approaches, but when captured well it makes it easy to identify what subject matter a text is about.

Ongoing Refinement

A further refinement by Joe Jeffrey (2002) was to extend the original Euclidean space into a hierarchy of nested subspaces, with a hierarchical distance function for matching of spectra across the subspaces. This makes it possible to represent hierarchies of information like Dewey Decimal, Library of Congress, and Medical Subject Headings, and also greatly reduces the work of creating and searching the space.

More recently we have worked out how to map a space of nested subspaces onto Euclidean space, thus simplifying distance calculations, and how to compress that space to a compactly represented Hamming space, thus saving space while regaining the speed advantages of the hierarchical space.

Latent semantic indexing, like many other techniques, is based on extracting the eigenvectors of the distribution of words across texts. Even the fastest means of extracting eigenvectors are of quadratic complexity, so we doubt these techniques can scale to the Web.

Judgment simulation does not require any quadratic algorithms, but it does require a lot more judgments than we can reasonably hope to pay for in the short term. So we have been analyzing the judgments of the 75,000+ Wikipedia editors, who have written over 2 million articles and classified them into over 200 thousand categories.

From Wikipedia to the Web

There are two kinds of judgments directly available in the Wikipedia data. One is the hierarchical structure of the categories, which reflects the judged relevance of each category to its parent. With these judgments we can create a Category Space. The other is the assignment of articles to categories, which reflects the judged relevance of articles to each category. With these judgments we can populate the space with articles. The final step is to position a vocabulary of words and phrases in the space. Then queries can be placed in the space based on their words and phrases, and nearby articles retrieved.

Rather than pay experts to develop the vocabulary, we are using the distribution of the words and phrases in Wikipedia text across the Wikipedia categories to get a position for each word and phrase in the category space. The statistical distribution of words and phrases is at best an indirect indication of their relevance to the various categories. But prior work by Kurz (2007, private communication) indicates that such statistics are effective proxies for direct relevance judgments, provided the categories are well chosen and the text is well categorized. We believe Wikipedia meets those criteria. Over time we can use paid and volunteer labor to improve the vocabulary as needed.

To scale beyond Wikipedia we will crawl the Web and place the pages we find into the category space at the centroid of their constituent words and phrases. The resulting space will be very large, with hundreds of thousands of dimensions and billions of data points. So we will be further pushing the state of the art with advanced techniques for efficient near-neighbor searches in high-dimensional space (Adoni & Indyk 2008).

References

1964, Ossorio. Classification Space Analysis. Defense Technical Information Center, Report AD0608034.

1966, Ossorio. Classification space: a multivariate procedure for automatic document indexing and retrieval. Multivariate Behavioral Research 1.

1976, Jeffrey. A new type of information retrieval system. Proceedings of the 14th ACM Southeast Regional Conference. Birmingham, Alabama.

1989, Ossorio & Kurtz. Automated classification of resolved galaxies. Proceedings of the Third International Workshop on Data Analysis in Astronomy, Plenum Press.

1990, Deerwester, Furnas, Landauer & Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41.

1991, Jeffrey. Expert Document Retrieval Via Semantic Measurement. Expert Systems With Applications, 2(4).

1992, Kurtz. Advice from the Oracle: Really Intelligent Information Retrieval. Adding Intelligence to Information Retrieval: The Case of Astronomy and Related Space Sciences, ed. Murtagh & Heck, Kluwer.

1993, Jeffrey. Judgment-Simulation Vector Spaces. Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Database, ComputerVision. Johns Hopkins University Press.

1999, Hofmann. Probabilistic Latent Semantic Indexing. Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval.

2000, Kurtz, Eichhorn, Accomazzi, Grant, Murray & Watson. The NASA Astrophysics Data System. Astronomy and Astrophysics Supplement Series 143:1.

2002, Jeffrey. Wide-spectrum information search engine. U.S. Patent 6,493,711.

2008, Adoni & Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM 51:1.