

Identifying Relationships within a Corpus

Myank Jain

Abstract

One of the pressing issues with information management today is actually finding relevant information among the vast amounts of data available. This paper discusses a methodology which extrapolates meaning from the contents of the documents themselves. This methodology is well suited for use as a search engine, as it focuses on the concepts contained within the documents, rather than on keywords found within the documents.

Introduction

“We are drowning in information and starving for knowledge.”

Rutherford D. Rogers, former Yale librarian

The 21st century is awash in a perfect storm of information. There is information all around us. It comes in every tongue known to man. It's also buried in an encrypted Tower of Babel. At the tail end of the 20th century, technology critic David Shenk took a look at the exponential expansion of this information glut and coined the phrase “data smog” to describe the phenomenon.

It's a good bet that this information has been digitized and can be discovered as web pages, e-mails, blogs, telephone call logs, message boards, survey responses and on countless hard drives and servers. If you can find it.

There's a lot of information out there. Today, Netcast reports, there are more than 182 million websites. Worldwide, a 2008 study by Radicati Group reveals, more than 210 billion e-mails are sent each day. Factor in countless online discussion groups and message boards, billions of domestic and international telephone calls, the explosive proliferation of blogs and the quantum grow of new one-to-many communications vehicle such as Twitter and other micro-blogging and social network applications, and a picture emerges of an information tsunami gathering strength minute by minute.

Information hierarchies, formal networks and organizational structures have been consigned to the dustbins of history. The information ecology is flattening. “It's natural enough to think of the growth of the blogosphere as a merely technical phenomenon,” *The New York Times* observes. “But it's also a profoundly human phenomenon, a way of expanding and, in some sense, reifying the ephemeral daily conversation that humans engage in. Every day the blogosphere captures a little more of the strange immediacy of the life that is passing before us. Think of it as the global thought bubble of a single voluble species.” It's a species that grows chattier by the day, which makes parsing the human conversation increasingly difficult.

Methodology

This proposed methodology leverages deep knowledge of human behavior, cognitive science and the analysis of unstructured data, resulting in the clustering of themes and concepts to show the most important threads, and surface patterns, and associations across the information network. This methodology also marries social and grammatical network analysis, natural language processing, token extraction and text analysis technologies to derive meaning from unstructured data.

The driving principle of this methodology is the simple fact that in any language there exist 'words'; and in any language, specific sets of words in context with each other define concepts. Context can be either derived or explicit. By utilizing current known statistical models to determine which concepts a document potentially intersects, a statistical representation of the document and all potential concept intersections can be created. These statistical representations are mapped into a theoretical n-dimensional space such as a Banach's space, where each axis is representative of a concept apparent in the system. Linkages between documents can be determined by the intersections with an axis, and the intensity of activity (the probability factor of the document discussing the concept) along a given axis.

Mapping a document into n-dimensional space in this fashion allows for analytics to be performed by a simple query. All of the processor-intensive calculations occur upon document digestion, and as such are incurred independently of any analysis. By condensing the single-concept axis into multi-concept axis, and adding these multi-concept axes into the n-dimensional construct representing the corpus, will efficiently identify complex relationships between documents. This approach facilitates seeding of the analysis with a concept, set of concepts, or a source document, and derives a network map of related concepts and associated documents.

The above-mentioned methodology analyzes a set of documents by transforming the document set into a network of documents (nodes) connected by terms (links). For example, two documents that have a common word will have a link in the network for this common term.

For large sets of documents the network graph will be complicated and dense. Furthermore, a single term linking two documents or even many documents does not imply anything in particular about whether those documents are related. This "Level 1" network graph, however, is the fundamental representation for this method.

Documents, of course, will frequently have more than one common term. Documents that have few common terms are most probably not related, meaning that they are probably not dealing with the same sets of concepts. Documents that have many common terms, however, have some probability of being related; it is probable that they are dealing with the same set of concepts. Methodology does not assert that there is meaning in relationships among documents that involve many common terms; it simply presents this information in an unbiased form to users so they can determine meaning for themselves.

A way to look at this process is to consider a layering scheme. The bottom layer is the "Level 1" network as described above. But, it is possible to create a "Level 2" layer where the links between documents represent two terms that are common. This graph will have the same number of documents but fewer links. This process of creating layers with links that represent more and more common terms can be continued until the very top layer is created where each link represents all of the words contained in the corpus of documents. This top layer has very little meaning because there are, most probably, no two documents in the set that contain all of the terms, but this layer provides a top level boundary or stop condition for creating the layers of information.

Making Sense

One can move from layer to layer exploring the term relationships, using them to establish themes or concepts among various groups of documents. Iteration is not linear, but cyclical, as the users develop meanings that cause them to return to earlier layers to refine term and document relationships.

At some layer (layer N) the term link consists of enough meaningful words to form a theme or concept. Because of the strong connection, this theme is likely to be a part of each document. That is, the documents are discussing at least one common idea. At this point, the user looks at other terms in these heavily related documents to see what other documents these new terms link to. This process pushes users down the stack of layers for the new terms and then refines their way back upwards again developing yet more themes that have meaning for them.

Because no information is excluded, and because no a-priori meaning is applied, the users have a reasonable probability of discovering meaning that is not otherwise obvious or part of their intention when they begins as well as uncovering a richer set of results for meaning that was part of their initial set of concepts. This activity exercises the transcendent term-interdependency because the methodology is not presuming any form of relevance or meaning on behalf of the user.

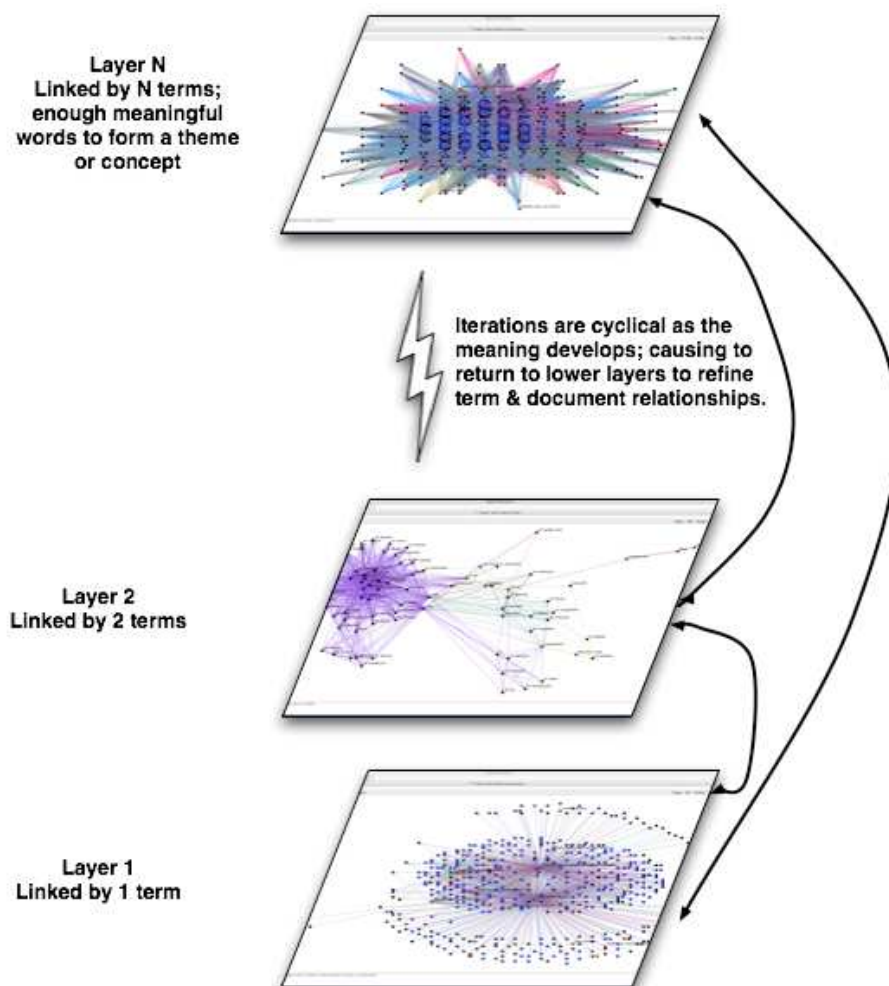


Figure 1- Methodology

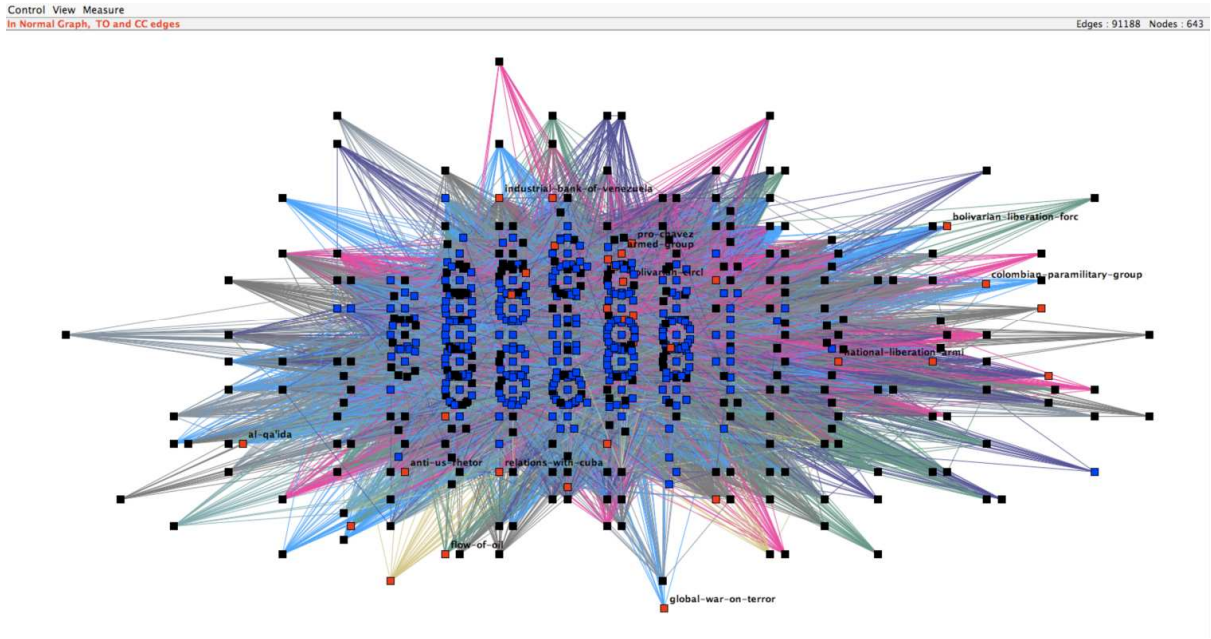


Figure 2 - Concept Map

Clusters are created showing documents and their term link density which is used to cluster documents with similar concepts.

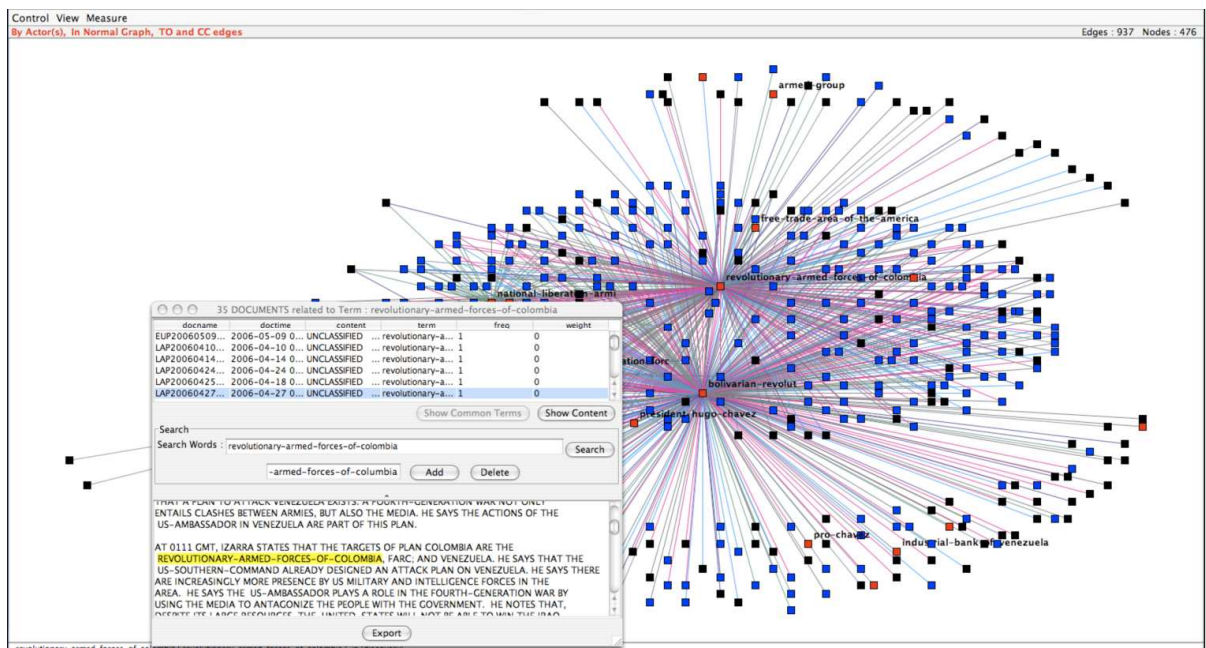


Figure 3 – Term Links

Layers m show the links that consist of m terms common to a pair of document. Term pairs are not pre-defined instead discovered in the data.

Conclusions

By using this methodology to generate disambiguation pages by extracting high occurrence concepts and presenting them to the user as choices a very powerful, and easy-to-use search technology can be developed. The primary query would be keyword driven, yet by providing the user with an intuitive approach to navigate from keywords into concepts via the disambiguation page, the transition from keyword to key-concept would be nearly invisible to the user. This approach fully leverages the concept-driven nature of the methodology, and allows the end user to explore search results in a familiar forum (text-only, keyword driven, statistically scored search results) while remaining efficient, and unencumbered by the mechanics of concept driven search.

References

Witten, I. H. and Frank, E. 2005 Data Mining: practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann.

Gloor, Peter A. 2006 Swarm Creativity, New York, NY: Oxford University Press.

Goldberg, D. E. 1989. Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.

Pyle, D. 1999. Data preparation for data mining. San Francisco, CA: Morgan Kaufmann.