Small worlds of concepts and other principles of semantic search

Stefan Bordag, Gerhard Heyer, Uwe Quasthoff

Leipzig University Computer Science Institute,
Natural Language Processing Department
Augustusplatz 10 / 11 D-04109 Leipzig
{bordag, heyer, quasthoff}@informatik.uni-leipzig.de

A combination of the strengths of both classic information retrieval with the distributed approach of P2P networks can avoid both their weaknesses: The organisation of document collections relevant for special communities allows both high coverage and quick access. We present a theoretical framework in which the semantic structure between words can be deduced from a document collection. This structural knowledge can then be used to connect document collections to communities based on their content.

Introduction

Comprising more than 3 billion documents, the WWW today contains the fastest growing collection of text. From the user's perspective, the traditional information retrieval (IR) problem of finding relevant documents that satisfy a particular description, has therefore become very urgent again..

A document is searched by locating all available (and indexed) documents that satisfy the search query description. In general, however, the set of retrieved documents does not consist of all and only the relevant documents, so that most likely the query needs to be repeated by a modified search. The actual process of generating a suitable description in a real life is frequently accelerated, or even made possible, by recurring to the semantic knowledge that is implicitly stored in the structure of human society or communities. In his famous article on small worlds [Milgram 1967], has succinctly drawn attention to the fact that passing an information from a source to some unknown addressee crucially depends on exploiting the semantic knowledge implicit in the description of the addressee to drastically reduce the search space for passing the message from one set of acquaintances to the next. In social reality, we apparently rely very heavily upon knowledge about which persons or communities are related to or engaged with which topics. It is the intention of this paper to explore how this implicit relation between the structure of a community and the structure of contents can be made more precise and exploited for the purposes of a semantic search.

Technically, advanced information retrieval systems like Gnutella, FreeNet, or NeuroNet, are based on the peer-to-peer (P2P) approach. This approach consists of a set of similar or compatible software agents that live on a network of connected com-

puters (paradigmatically the internet). Each software agent can likewise act as client and as server, and is accordingly called a servent. Each servent comprises a data base with the IP-address of its neighbours that host servants belonging to the same system. At present, P2P systems are mainly used as filesharing systems; semantic principles for processing queries as sketched above have not been pursued so far.

Words, Document Collections, Link Structure, and Communities

If we define a community as a group of people sharing some common interest, there should be a collection of documents which is of interest to all of them. If, moreover, the documents are available in the web, some of them will be linked via hyperlinks. Many of these documents will contain some words or phrases which are specific to the interest of this community. Any meaningful classification of words can help to classify documents and hence, to identify communities.

However, any such classification has to be deal with polysemy and ambiguity. These properties are inherited from natural language, but also appear in documents (because a document can adress more than one interest) and members of a community (because they can have multiple interests).

In contrast to this common difficulties we also find the following common structural feature: Members of communities, the hyperlink structure of the web [Barabasi 2000] and words according to their semantics [Ferrero 2001] all form so-called small worlds [Strogatz 1998]. These similarities are used to describe a framework how to extend a classification of words to find communities in the web.

Semantic Structures

To motivate the classification of words we start with structuralist semantics and put these relations in a statistical context.

Structuralist Semantics

Our main thesis is that the structure of a content can be derived from a set of documents produced and exchanged within a community. Following the famous Swiss linguist Ferdinand de Saussure, meaning (and other notions of linguistic description) can be defined solely by reference to the structural relations existing amongst the words of a language [Saussure 1916]. Syntagmatic and paradigmatic relations between words constitute the basis for such relations.

Examples of syntagmatic relations typically include dependencies between nouns and verbs, enumerations or the compounding of nouns and nouns, and head-modifier constructions based on adjectives and nouns or nouns and nouns. Paradigmatic relations vary depending on the measure of similarity presumed. On the syntactic level, paradigmatic relations typically comprise distribution classes for the main syntactic categories. On the semantic level, paradigmatic relations range from semantic fields

to well defined logical relations such as hyponymy, co-hyponymy, hyperonyms, synonyms and antonyms.

Let L be a natural language and W be the set of full form words of this language.

Then any sentence S of L represents a sequence of word forms $S = \{w_1,...,w_i,...,w_n\}$ with all $w_k \in W$.

By the *context* of a word form $w_i \in S$ we mean a subset of all word forms occurring in S suitably chosen:

$$K_S(w_i) \subset \{w_1,...,w_i,...,w_n\}.$$

Usually this subset will contain the meaningful words according to some statistical measure to be defined later. Similarily, the exact meaning of the *most*-operator $M\,M$ and the set similarity ~ used below have to be defined.

Abstract *syntagmatic* and *paradigmatic* relations of two word forms w_i and w_k can now be defined as follows:

Common joint appearance of two word forms w_i and w_k defines the abstract *syntagmatic relation SYN*: Two word forms w_i and w_k are related syntagmatically iff most of the contexts of w_i contain the word w_k:

$$(M\ S : w_i \in S)(w_k \in K_S(w_i) \rightarrow SYN(w_i, w_k))$$

Joint context shared by two word forms w_i und w_k defines the abstract *paradigmatic relation*: Two word forms w_i and w_k are related iff they usually appear within similar contexts:

$$(M\ S : w_i \in S\ \exists K(w_k) : K_S(w_i) \sim K(w_k)) \rightarrow PARA(w_i, w).$$

Co-occurrences

Some words co-occur with certain other words with a significantly higher probability and this co-occurrence is semantically indicative. We call the occurrence of two or more words within a well-defined unit of information (sentence, document) a *collocation*. For the selection of meaningful and significant collocations, an adequate collocation measure has to be defined.

Let a, b be the number of sentences containing A and B, k be the number of sentences containing both A and B, and n be the total number of sentences.

Our significance measure calculates the probability of joint occurrence of rare events. The results of this measure are similar to the *log-likelihood*-measure:

Let $x = ab/n$ and define:

$$sig(A,B) = \frac{-\log\left(1 - e^{-x}\sum_{i=0}^{k-1}\frac{1}{i!}\cdot x^{i}\right)}{\log n}$$

For $2x < k$, we get the following approximation, which is much easier to calculate:

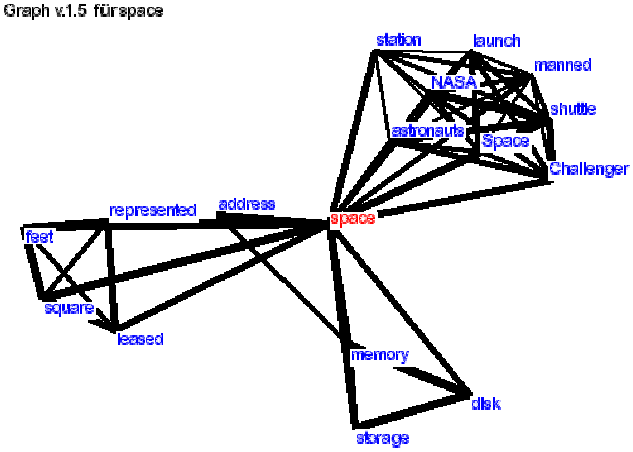
$$sig(A,B) = \frac{(x - k\log x + \log k!)}{\log n}$$

In general, this measure yields semantically acceptable collocation sets for values above an empirically determined positive threshold. Hence, we can use this measure to select the relevant words in a sentence and to determine the context of a word as described in the above section.

Example: *space*

Fig. 1 shows the collocations of the word space. Two words are connected if they are collocations of each other. The graph is drawn useing *simulated annealing* (see [Davidson 1996]). Line thickness represents the significance of the collocation. The resulting picture represents semantic connectedness surprisingly well. In Fig. 1 we find three different meanings depicted: real estate, computer hardware, and astronautics.

Fig. 1. Collocation Graph for space



The connection between *address* and *memory* results from the fact that address is another polysemous concept. This kind of visualization technique can also be applied to knowledge management techniques, esp. the generation and visualization of Topic Maps [Böhm 2002].

Small Worlds

It is worth to take a closer look at the graph which is implicitly defined by the co-occurrence measure. Let $G = (N, C)$ be a graph with N as the set of nodes and C the set of edges between the nodes. The nodes are labelled by different words of our language, we only consider words having collocations as described above. Two nodes are connected with each other if they are collocations. The resulting graph has several important properties that will be briefly described:

- The graph is *nearly fully connected*, i.e. for two randomly chosen nodes there is a high probability that they are connected in this graph.
- It is *sparse*. Usually the number of edges is approximately only one order higher than the number of nodes.
- It has the *small world* property.

A graph having the *small world* property is a graph which has both short average path lengths like a random graph, as well as high local clustering coefficients like a regular graph. First formalizations and the explicit differentiation of the small world property of graphs contrasting the traditional extremes, the regular and random graphs, have been introduced by [Strogatz 1998]. Further work can be found at [Kleinberg 2000] and [Ferrero 2001]. In short, a *random graph* is a graph where nodes are connected randomly, and which has a given degree distribution of the nodes which can, for example, be power-law or exponential. A *regular graph* is a graph where all nodes have the same number of connections to other nodes.

The *path length* between two nodes in the graph is denoted by $d_G(i, j)$ with $i, j \in N$ and measures how many connections part the two given nodes at least. The average path length d_G over the graph is then calculated as the arithmetical mean over all possible distances:

$$d_\Omega = \frac{1}{|N|} \sum_{i=|N|}^{1} \left(\frac{1}{i} \sum_{j=1}^{i} d_G(i, j) \right)$$

The *clustering coefficient* c_i of a node i compares the number of connections T_{Γ_i} between the neighbours Γ_i of the given node with the number of possible connections:

$$c_i = \frac{2T_{\Gamma_i}}{|\Gamma_i| \cdot (|\Gamma_i| - 1)}$$

For the whole graph, the clustering coefficient c_G can then be calculated as a mean over the clustering coefficients of each node in the graph:

$$c_\Omega = \frac{1}{|N|} \sum_{i=1}^{|N|} c_i$$

The clustering coefficient thus measures the probability that two nodes are connected with each other if they are connected to a common third node.

A comparison of the two graphs indicates that a random graph will always have much shorter path lengths than a regular graph if both have the same number of nodes and connections. If graphs are sparse graphs, then with growing graph size the path length will grow linearly for the regular graph and only logarithmically for the random graph because of the many possible shortcuts throughout the entire graph. On the other hand, the clustering coefficient will always be very low for the random graph as opposed to the regular graph.

The small-world property has originally been used to explain why certain random graph based models resulted in wrong predictions about phenomena like disease spreading. In these cases, the high neighborhood clustering together with the short path lengths lead to a much more efficient model of graph formation when compared to random graphs. It is the intention of the present paper to draw attention to the fact that similar phenomena of small world formation can be detected with respect to the conceptual space of internet communities, and that this fact might be exploited for implementing more efficient, semantically based search strategies.

Disambiguation

Based on the co-occurrences of word forms and the small-world property of their collocations graph, an approach to solve the polysemy problem has been introduced by [Bordag 2002b]. Applications include improved text classification methods, improvements in Word Sense Disambiguation algorithms, better query expansion, intelligent spell checking and more.

The algorithm is based on two assumptions: first, words in the graph cluster semantically and, second, any three given words are unambiguous (there are only few cases where this does not hold). If three words are semantically homogenous, they are located in the same cluster of the graph. The intersection of their direct neighbours will not be empty, and they will be semantically homogeneous as well. After generating an amount of such triplets (always including the input word), their neighbour-intersections are clustered with hierarchical agglomerative clustering.

As a result, for a given word one or more sets of semantically homogeneous words are found along with a set of words which are either semantically unrelated to the input word (although they are co-occurring with it), or whose statistical count is not high enough to make a reliable decision. Problems occur when a corpus is unbalanced with respect to certain sub-languages where certain usage contexts of a word are missing.

Graph based Automatic Semantic Convergence (ASC)

Combining the algorithm sketched above with known methods and linguistic data resources, we introduce a first framework for a semantically based search, aiming at a system by which a variety of textual information can be processed in a fully unsupervised manner.

Instead of keeping a central index of the content of a WEB network, we propose agents analogous to the common P2P agents. But unlike agents that simply collect a set of IP addresses of neighbours in the network and broadcast queries to them, we propose *dynamic* agents. These dynamic agents should be able to decide intelligently what to do with a query, either answer it based on the documents available to the agent, or forward it to an agent which would better satisfy the query.

Abstract definition of the convergence framework

Basically, an ASC-agent consists of a set of documents, the related semantic knowledge database, and a set of IP addresses of neighbours where other agents with the same interface reside. The lifecycle of an agent consists of periodic comparisons of its knowledge database with those of its neighbours. After such a comparison it is decided which of its neighbours has the least semantically fitting content, and the worst ones are dropped in favor of better ones. How exactly the new neighbours are chosen, or how many are dropped, and many other options can be left as implementation parameters and may easily differ from one agent to another to tune agents to specific needs. Nevertheless, an agent should be required to reserve a small fraction of its connections to agents that have a large number of outgoing links. This corresponds to small worlds where a subclass differs from others in that it has only a few hub-like nodes that connect to large parts of the network at once. It can also be left open whether the connections should be symmetrical or directed.

We begin by defining an agent A as a sixtuple $A = (\Gamma, \Delta, \phi, \Theta, c, a)$ where

Γ	set of neighbours
Δ	set of owned content (i.e. a set of Documents)
$\phi : (A_1, A_2) \to [0..1]$	node similarity operation $[0..1]$
$\Theta : A \to A'$	convergence operation
c	connectivity threshold
a	activity threshold

Both the node similarity operation ϕ and the convergence operation Θ are left unspecified although some possibilities will be discussed.

For the similarity operation, any traditional document comparison model can be used but there are two important constraints: First, the set of all documents is always unknown (even its size) and, second, the set of terms used in these documents is unknown as well. That means that for example the Vector Space Model will have to be built from scratch from Γ_{A_1} and Γ_{A_2} dynamically for each comparison.

The convergence operation is more intricate as it might have great impact on the overall behavior of the network. Two main possibilities emerge at this point. The first has already been mentioned:

- Compare own content to content of neighbours
- Drop some of the worst ones
- Replace them with better ones
- Keep a fraction of connections to highly connected nodes, no matter how bad they fit semantically

The second one is more radical:

- Use Text classification algorithms to build clusters from contents of neighbours including own content
- Replace the worst cluster with neighbours of the best cluster and a fraction of random connections with highly connected nodes

In the second case, a node might enter a state where it would have to remove itself along with the worst cluster. It would then have to start again completely randomly at a different place in the network to perform better than the first time. Improvements can be imagined by making use of links found in the own document set and trying to use them to find relevant agents directly, speeding up the semantic convergence.

Another important aspect of such document-representing agents is that they can be inherently ambigue themselves. They will certainly contain documents from more than one topic. This means that they will have to be able to handle this properly. Here is where the idea of above described disambiguation algorithm can be reused. As it is based on a very similar construct, a sparse graph having clusters, it should be possible to alter it in order to fit it to this task. As such it will provide a robust unsupervised clustering of the topics in the local document collection.

Research on the small-world properties of graphs indicates that the above network is most likely to converge to clusters of agents with similar content, exhibiting the small-world property. From that follows that queries, once they are handed over to an agent in such a cluster, are either answered immediately or by handing them just one or two steps further to the best possible agent without broadcast. In case that a query begins with a completely unfitting agent, the agent decides to hand it over to the agent which has the highest connectivity trusting that this new agent will have a connection to a distant agent which might be more fitting than anything it had itself. The short path lengths in this network will have the effect that a search query, although never broadcasted, will not have to travel far until it reaches its destination.

Conclusion

By our approach, the role of the classic P2P agent changes in that it is not only a mechanic collection of links and files. The network comprised by the semantic agents sketched above evolves on the basis of the content on which they ‘reside’. In a sense, the agents are not only aware of where they are but also of what they represent. It is important that all components of such agents are well known and robust algorithms and methods. The most important aspect, however, is that a user who decides to participate in a network by installing an agent has to do nothing except pointing the agent to the documents it should represent. This is in contrast to current WEB projects like semantic web where users are encouraged to improve the quality of the WEB and its services by providing manually created metadata for their data.

References

[Adar 2000] E. Adar, B. Hubermann. Freeriding on Gnutella. Firstmonday 5(10), 2000

[Barabasi 2000] A.L. Barabasi et al . Scale-free characteristics of random networks: the topology of the World-wide web, Physica A (281)70-77, 2000

[Bolloba 1985] B. Bolloba. Random Graphs, Academic Press, London, 1985

[Bordag 2002b] Stefan Bordag, Sentence Co-occurrences as Small-World Graphs: A solution to Automatic Lexical Disambiguation, A. Gelbukh (Ed.): CICLing 2003, LNCS 2588, pp. 329-332, Springer-Verlag Berlin Heidelberg, 2003

[Bordag 2002a] S. Bordag. Vererbungsalgorithmen von semantischen Eigenschaften auf Assoziationsgraphen und deren Nutzung zur Klassifikation von natürlichsprachlichen Daten, Diplomarbeit, Universität Leipzig, Institut für Mathematik und Informatik, 2002

[Böhm 2002] K.Böhm, G. Heyer, U.Quasthoff, Chr. Wolff. Topic Map Generation using Text Mining, Journal of Universal Computer Science, Vol.8, Issue 6, http://www.jucs.org/jucs_8_6, 2002

[Clarke 2000] I. Clarke, O. Sandberg, B. Wiley, T. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. ICSI Workshop on Design Issues in Anonymity and Unobservability, Berkeley, CA, 2000

[Davidson 1996] R. Davidson, D. Harel. Drawing graphs nicely using simulated annealing. ACM Transactions on Graphics. vol. 15, num. = 4, pp. 301-331. 1996

[Deo 2001] N. Deo, P. Gupta. World Wide web: a Graph Theoretic Approach. Technical Report CS TR-01-001, University of Central Florida, Orlando Fl. USA, 2001

[Ferrero 2001] R. Ferrero i Cancho, R. V. Solé. The Small-World of Human Language. (http://www.santafe.edu/sfi/publications/), 2001

[Gnutella] Gnutella. www.gnutella.com. 2002

[GRACE] GRACE http://www.ub.uni-stuttgart.de/grace/ 2002

[Heyer 2000] G. Heyer, U.Quasthoff, C.Wolff. Aiding Web Searches by Statistical Classification Tools. Knorz, Gerhard; Kuhlen, Rainer (edd.). Informationskompetenz - Basiskompetenz in der Informationsgesellschaft. Proc. 7. Intern. Symposium f. Informationswissenschaft, ISI 2000, Darmstadt. Konstanz: UVK, 163-177, 2000

[Heyer 2001] G. Heyer, U.Quasthoff, T.Wittig, C.Wolff. Learning Relations Using Collocations, A. Maedche, S. Staab, C. Nedellec and E. Hovy, (eds.), Proc. IJCAI Workshop on Ontology Learning, Seattle/ WA, 2001

[Heyer 2002a] G. Heyer, U.Quasthoff, C.Wolff. Automatic Analysis of Large Text Corpora - A Contribution to Structuring WEB Ccommunities, in: H.Unger, Th. Böhme (Hrsg.), Proceedings I2CS - 2002, Advanced Lecture Notes in Computer Science, Springer: Berlin, Heidelberg, New York 2002

[Heyer 2002b] G. Heyer, U.Quasthoff, C.Wolff. Knowledge Extraction from Text: Using Filters on Collocation Sets, LREC-2002 IICS 2002, Lectures Notes in Computer Science 2346, 153 - 162, Springer, 2002

[Joseph 2001a] S. Joseph: NeuroGrid - Freenet Simulation Results.

[Joseph 2001b] S. Joseph: NeuroGrid White Paper.

[Kleinberg 2000] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing, 2000

[Lechner 2002] U. Lechner: Peer to Peer beyond Filesharing. In H.Unger, T. Böhme (Hrsg.)

[Lifantsev 2000] M. Lifantsev. Voting Model for Ranking Web Pages, In Peter Graham and Muthucumaru Maheswaran, editors, Proceedings of the International Conference on Internet Computing (Las Vegas, Nevada, U.S.A.), CSREA Press, pages 143-148, Las Vegas, 2000

[Lifantsev and Chiueh 2002] M.Lifantsev and T.Chiueh. I/O-Conscious Data Preparation for Large-Scale Web Search Engines. Proceedings of 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China, Morgan Kaufmann, Hong Kong, 2002

[Milgram 1967] S. Milgram. The small world problem. Psychology Today 2, pp. 60-67, 1967.

[Neurogrid] www.neurogrid.com

[Newman 2000] M.E.J. Newman. Models of the Small World, 2000

[OpenGriD] Open GriD http://www.cs.sunysb.edu/~maxim/OpenGRiD/ 2002
[Ritter 2002] J. Ritter. Why Gnutella can't scale. No, really. www.nearlydeaf.8m.com/ygnutellwnwrk.html. 2002
[Sanderson 1996] M. Sanderson (1996): Word Sense Disambiguation and Information Retrieval; Proceedings of the 17th ACM SIGIR Conference, pp. 142-151
[Saussure 1916] Saussure, F. de Saussure, Cours de linguistique générale, 1916
[Schmidt 1999] F. Schmidt. Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung, Diplomarbeit, Universität Leipzig, 1999
[Manning & Schütze 1999] C. D. Manning & H. Schütze. Foundations of statistical natural language processing, 1999.
[Sebastiani 2001] F. Sebastiani. Machine Learning in Automated Text Categorization, 2001.
[Singla 2002] A. Singla, C. Rohrs. Ultrapeers: Another Step towards gnutella scalability. Lime Wire LLC, Working Draft, 2002. www.limewire.com/developrer/Ultrapeers.html
[Steyvers & Tenenbaum 2002] M. Steyvers, J. B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. M. Steyvers, J. B. Tenenbaum, Cognitive Science, 2002
[Strogatz 1998] D. J. Watts, S.H. Strogatz. Collective dynamics of 'small-world' networks, Nature 393:440-442, 1998.
www.firstmonday.org
www.neurogrid.net/ng-simulation.html (2001).
www.neurogrid.net/WhitePaper0_3.html (2001).