Calculating Communities by Internet Link Analysis

Gerhard Heyer, Uwe Quasthoff

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

Collocation analysis finds semantic associations of concepts using large text corpora. If the same procedure is applied to sets of outgoing links of web pages, we can find semantically related web domains. The structure of the semantic clusters shows all properties of small worlds. The algorithm is known to work for large parts of the web like the German internet. As a sample application we present a surf guide for the German web.

Introduction

When analyzing the internet, the term *community* generally refers to a collection of web pages that offer content to one and the same topic and contain links of each other. It is well known that within internet communities there exists a characteristic link structure that can be measured and described by graph theoretical means [Gibson 1998], [Barabasi 2000]. Among the key aspects to be taken into account describing communities, we briefly rehearse the following (cf. [Brinkmann 2003]):

- **Structure:** Prototypical structures comprise centralistic structures, generally known as authorities and hubs, and so-called webrings [Deo 2001]. In reality, however, an internet community will contain a *variety* of link structures.
- **Non-exclusiveness:** Participants in a community can be members in another community, too. Membership to a community must not be exclusive.
- **Hierarchy:** Communities can themselves be part of larger communities. In general, a community that is contained by a larger community can be considered a specialisation of the more general one.
- **Communication:** Communities can be related, i.e. there may be communication links between communities.

It is also generally accepeted, that an algorithm for detecting communities must fulfill the following requirements:

- **Stability:** The algorithm should yield nearly the same results when fed with the slightly disturbed data. In particular, the choice of a starting point for calculating the community should have no effect on the result.
- **Performance:** In order to efficiently calculate community structures, the performance of the algorithm should have a complexity between linear and

quadratic complexity. Otherwise communities could be computed by means of clustering algorithms known to have cubic complexity.

In what follows, we present an approach to calculating internet communities based on a natural language processing technology for calculating semantic networks of words. The basic idea is that if we are interested in the characteristic concepts of a certain subject area, we can take some known concepts of this subject area and look for concepts co-occurring significantly often with those starting concepts. This co-occurrence can be measured for texts using a window size of one sentence. Here we want to apply the same procedure to URLs: Assuming that URLs often mentioned on the same web page belong to the same subject area, we want to generate a cluster of semantically related URLs.
It turns out that the algorithms developed for the semantic analysis of natural language yield promising results for the semantic analysis of the internet.

Background: Collocations

Some words co-occur with certain other words with a significantly higher probability and this co-occurrence usually turns out to be semantically indicative. We call the significant cooccurrence of two (or more) words within a sentence a *collocation*. For the selection of meaningful and significant collocations, the following collocation measure has been defined (cf. [Quasthoff 2002]).

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.

The significance measure uses a Poisson distribution and calculates the probability of the 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:

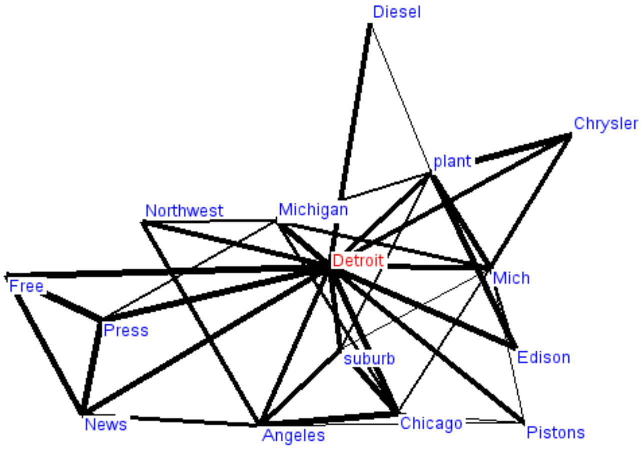
$$sig(A,B) = \frac{\left(x - k \log x + \log k!\right)}{\log n}$$

In general, this measure yields semantically acceptable collocations for values above an empirically determined threshold.

Example: *Detroit*

Fig. 1 shows the collocations of the word *Detroit*. 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. In Fig. 1 we find different aspects of *Detroit*: Mainly other cities related to *Detroit* and names of organizations based in *Detroit*.

Fig. 1. Collocation Graph for *Detroit*



Link Analysis

The following link analysis was performed for a large part of the German web, i.e. for “.de”-domains.

The analysis of URLs follows the collocation analysis described above as close as possible. Analyzing sentence collocations, we analyse a large corpus of monolingual text and look for words (or concepts) occurring significantly often together within sentences. Both, the number of sentences and the number of different words are usually in the range of 10^6. A typical sentence contains about 10 words.

In the case of link analysis we will get numbers in the same range if we consider the following units for our analysis:

1. URLs found in link targets replace words. Web pages replace sentences. For a given web page, the corresponding ‘sentence’ contains sets of link targets found on this page. Their order is irrelevant.
2. Two URLs are said to co-occur in such a sentence if there are links to both of them on the original web page generating the sentence.

3. We only consider top-level-domains. Hence, both www.xyz.de/index.html and www.xyz.de/programm/2004.html are mapped to the same URL www.xyz.de. Dubletts are removed from the sentences.

The number of sentences generated this way again is in the range of 10^6. The number of different words also has the desired size of 10^6. The reason for these numbers, relatively small compared to the actual number of web pages, is as follows. First, only part of the German web was crawled. Second, many web pages do not contain links to other URLs. This reduces the number of sentences about a factor of 10. Let us consider the individual steps of the analysis in detail.

Step 1: Sets of Links as Sentences

The following box gives two sample sentences as stored in the database. For inspection reasons the fist entry is the URL of the page containing the links given in the second part:

```
http://www.jazzdimensions.de/interviews/portraits/craig_schoedler.html
    www.craigschoedler.com   www.atomz.com   www.phonoclub.de
http://www.google.de/appliance/index.html   www.reuters.com   www.infoworld.com
    www.ecommercetimes.com   services.google.com
```

Step2: Collocations of Links

In the analysis of natural language text, collocations have proved to be useful to discover pairs of words connected by a semantic relation. The software available for calculating collocations between words works for large texts of 10^9 running words (tokens) containing 10^7 different words (types).

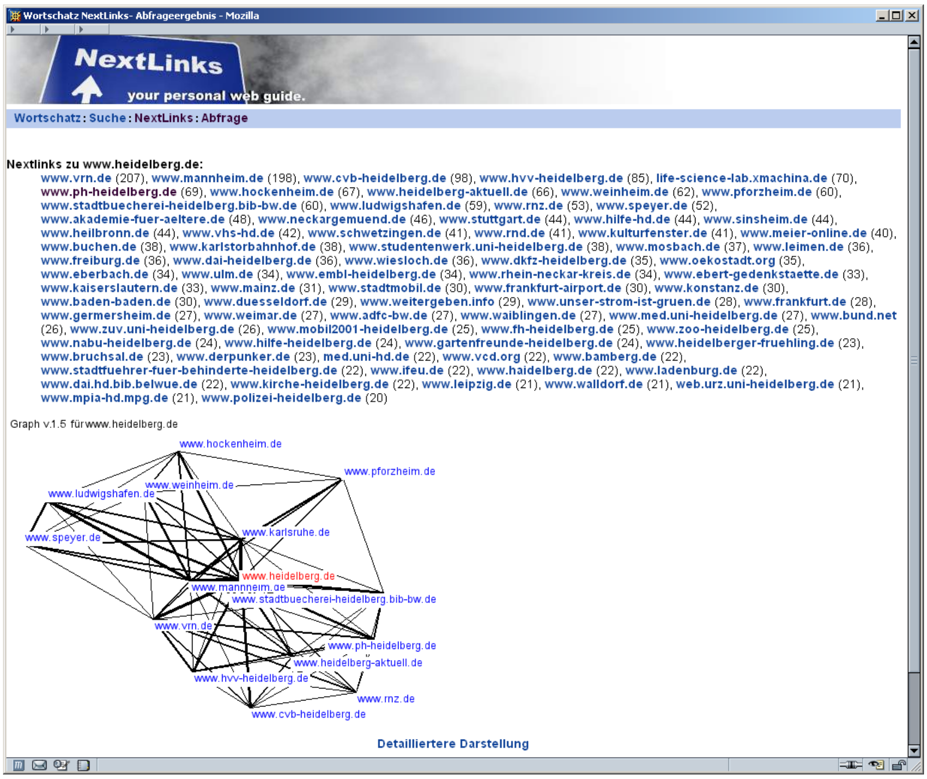
Now, the same procedure is applied on sets of URLs instead of sentences of words.

To calculate the similarity of two URLs A and B using the formula given above, now let a, b be the number of incoming links for A and B, k be the number of pages containing both links to A and B, and n be the total number of URLs considered.

Due to the data structure we can apply exactly the same algorithms and software. Sample results are given in the next section.

Comparison of URLs and Words

Figure 2 shows the most similar URLs for *www.heidelberg.de* both in a list ordered by similarity and drawn using simulated annealing as described above.


Wortschatz NextLinks- Abfrageergebnis - Mozilla
NextLinks
your personal web guide.
Wortschatz : Suche : NextLinks : Abfrage
Detailliertere Darstellung

Nextlinks zu www.heidelberg.de:

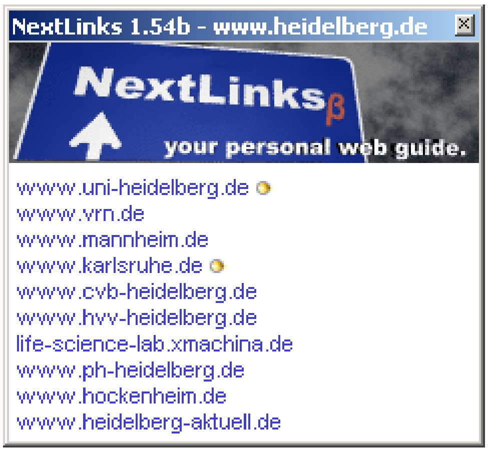
www.vrn.de (207), www.mannheim.de (198), www.cvb-heidelberg.de (98), www.hvv-heidelberg.de (85), life-science-lab.xmachina.de (70), www.ph-heidelberg.de (69), www.hockenheim.de (67), www.heidelberg-aktuell.de (66), www.weinheim.de (62), www.pforzheim.de (60), www.stadtbuecherei-heidelberg.bib-bw.de (60), www.ludwigshafen.de (59), www.rnz.de (53), www.speyer.de (52), www.akademie-fuer-aeltere.de (48), www.neckargemuend.de (46), www.stuttgart.de (44), www.hilfe-hd.de (44), www.sinsheim.de (44), www.heilbronn.de (44), www.vhs-hd.de (42), www.schwetzingen.de (41), www.rnd.de (41), www.kulturfenster.de (41), www.meier-online.de (40), www.buchen.de (38), www.karlstorbahnhof.de (38), www.studentenwerk.uni-heidelberg.de (38), www.mosbach.de (37), www.leimen.de (36), www.freiburg.de (36), www.dai-heidelberg.de (36), www.wiesloch.de (36), www.dkfz-heidelberg.de (35), www.oekostadt.org (35), www.eberbach.de (34), www.ulm.de (34), www.embl-heidelberg.de (34), www.rhein-neckar-kreis.de (34), www.ebert-gedenkstaette.de (33), www.kaiserslautern.de (33), www.mainz.de (31), www.stadtmobil.de (30), www.frankfurt-airport.de (30), www.konstanz.de (30), www.baden-baden.de (30), www.duesseldorf.de (29), www.weitergeben.info (29), www.unser-strom-ist-gruen.de (28), www.frankfurt.de (28), www.germersheim.de (27), www.weimar.de (27), www.adfc-bw.de (27), www.waiblingen.de (27), www.med.uni-heidelberg.de (27), www.bund.net (26), www.zuv.uni-heidelberg.de (26), www.mobil2001-heidelberg.de (25), www.fh-heidelberg.de (25), www.zoo-heidelberg.de (25), www.nabu-heidelberg.de (24), www.hilfe-heidelberg.de (24), www.gartenfreunde-heidelberg.de (24), www.heidelberger-fruehling.de (23), www.bruchsal.de (23), www.derpunker.de (23), med.uni-hd.de (22), www.vcd.org (22), www.bamberg.de (22), www.stadtfuehrer-fuer-behinderte-heidelberg.de (22), www.ifeu.de (22), www.haidelberg.de (22), www.ladenburg.de (22), www.dai.hd.bib.belwue.de (22), www.kirche-heidelberg.de (22), www.leipzig.de (21), www.walldorf.de (21), web.urz.uni-heidelberg.de (21), www.mpia-hd.mpg.de (21), www.polizei-heidelberg.de (20)

Graph v.1.5 für www.heidelberg.de

The graph of the URL *www.heidelberg.de* shows strong similiarities to the graph of the word *Detroit*: Both the word and the URL of the cities are connected to the same type of objects. On one hand, they are connected to other cities nearby and/or of the same size, on the other hand, they are connected to organisations located in the city. Hence, the identical ananalysis of both text and links represent the underlying semantics in the same way. This might lead to the conclusion that human authors use links the same way as they use words: In the typical case, both words and links are choosen carefully according to their semantic content. In the case of proper names, both the word and the URL are used to denote the same object. What the algorithms find out mihgt be the relations between those objects, regardless of their representation.

Application: NextLinks

In order to test the results for user acceptence, we implemented a *surf guide* called NextLinks which displays the top-10 similar URLs for the URL found in the browser window, see figure 3 for *www.heidelberg .de*.



www.uni-heidelberg.de ●
www.vrn.de
www.mannheim.de
www.karlsruhe.de ●
www.cvb-heidelberg.de
www.hvv-heidelberg.de
life-science-lab.xmachina.de
www.ph-heidelberg.de
www.hockenheim.de
www.heidelberg-aktuell.de

At the moment, the following data are used:

Number of URLs crawled	980.751
Number of different domains found	886.107
Number of domains with similar domains found	351.033

Table 1: Amount of data for domain similarity

NextLinks is available from *http://wortschatz.informatik.uni-leipzig.de/nextlinks/*

Further steps

To get deeper insights into the link structure of the web we need more data. The data used here were crawled with nedlib [NEDLIB]. The next dataset will be crawled by a distributed system having many clients for crawling and link extraction.
The similarity between links and words shown for cities can be carried further if one analyses the strings used to name domains and subdomains. Here we can find even more relations between URLs and words.

References

[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

[Brinkmeier 2003] M. Brinkmeier. Communities in Graphs, in: Böhme, Th., Heyer, G., Unger, H. (eds.), Innovative Internet Community Systems, Proceedings of the Third International Workshop I2CS 2003, Leipzig, 20-35, Springer: Berlin, Heidelberg, New York 2003

[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

[Gibson 1998] D.Gibson, J.Kleinberg, P.Raghavan. Inferring Web Communities from Link Topology. in Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, Pennsylvania, pp. 225-234, 1998

[NEDLIB] NEDLIB Harvester, *http://www.csc.fi/sovellus/nedlib/*

[Quasthoff, U. 2002] U.Quasthoff, Chr. Wolff. The Poisson Collocation Measure and its Applications,.in: Proc. Second International Workshop on Computational Approaches to Collocations, Wien, Juli 2002.