Visibility Analysis on the Web Using Co-visibilities and Semantic Networks

Peter Kiefer, Klaus Stein, and Christoph Schlieder

Laboratory for Semantic Information Processing
Otto-Friedrich-University Bamberg, Germany
{peter.kiefer, klaus.stein, christoph.schlieder}@wiai.uni-bamberg.de

Abstract. Monitoring public attention for a topic is of interest for many target groups like social scientists or public relations. Several examples demonstrate how public attention caused by real-world events is accompanied by an accordant visibility of topics on the web. It is shown that the hitcount values of a search engine we use as initial visibility values have to be adjusted by taking the semantic relations between topics into account. We model these relations using semantic networks and present an algorithm based on Spreading Activation that adjusts the initial visibilities. The concept of co-visibility between topics is integrated to obtain an algorithm that mostly complies with an intuitive view on visibilities. The reliability of search engine hitcounts is discussed.

1 Introduction

Social scientists have invested much effort in manually analyzing daily news while trying to monitor public awareness for certain topics (see e. g. [1]). Especially in nowadays information society, the topics that are visible in public discussions across different kinds of media tend to change rapidly. It becomes increasingly important for organizations to be present in the minds of people and to evaluate public relations activities [2], be it a company competing for customers' attention or a non-profit organization trying to arouse public awareness for their concerns (see also work on attention economies, e. g. [3]). The undoubted primacy of the internet raises the question whether public visibility of topics goes along with an accordant visibility of these topics on the web. If such a correlation between real world events and online visibility exists, monitoring topics on the web could give an important indicator for the target groups mentioned above.

In this paper, we aim at providing methods to support the monitoring of the visibility of topics on the internet. We do not deal with topic detection, but assume a user who previously knows the topics of interest. We thereby take a quite broad view of what is regarded as a topic: anything that can draw public attention on itself (and is expressible by some kind of search term), ranging from typical discussion group topics like 'climate policy' to persons like 'George Bush' or even something basic like 'Christmas'. We propose a simple way to measure the visibility of topics, based on hitcount values of a search engine, present examples

M. Ackermann et al. (Eds.): Semantics, Web, and Mining 2005, LNAI 4289, pp. 34–50, 2006.

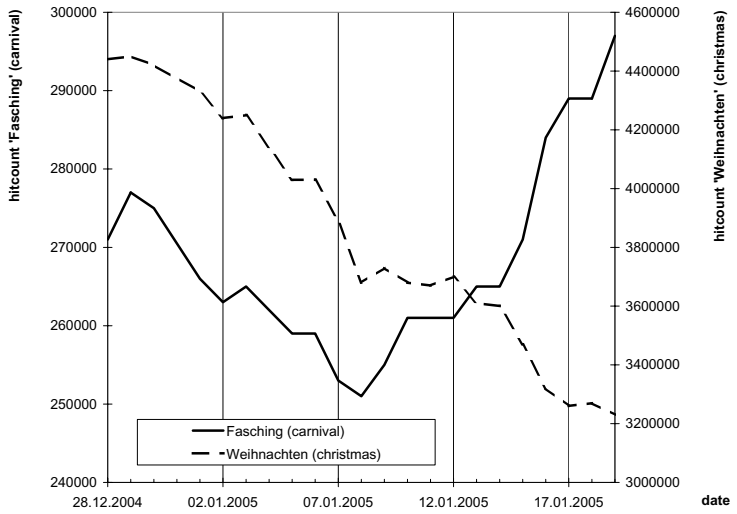

Fig. 1. Estimated hitcount values (Google) for 'Weihnachten' (Christmas) and 'Fasching' (Carnival) in time

indicating that real world events actually do have an impact on visibility on the web and introduce the concept of topic co-visibility (section 2).

It is often not sufficient to monitor just a single topic, rather several semantically related topics need to be observed simultaneously. We show how to correct our initial visibility values by adding knowledge about the semantical relations between topics (section 3). In this context, we contribute a new algorithm based on Spreading Activation (section 4).

In section 5 we report on the experiences we made concerning the reliability of the hitcounts of the search engine we used for our case studies. At last (section 6) we give a short summary and a view on current research.

2 Visibility and Co-visibility

2.1 Visibility

Our first objective was to find an appropriate measure for the visibility of a topic in internet communication processes. However, possible measures depend on the communication process analyzed, for instance messages in a newsgroup should be treated differently than a collection of documents without link structure. We define the visibility of a given topic by vis(top) = hitcount("top") with hitcount("top") being the number of pages found on the search term "top" by a given search engine.[1]

[1] For all examples given in this paper we used the estimated hitcount values of the Google Web API (`http://www.google.com/apis/`). Note that hitcount values from search engines (especially from Google) are usually estimated and not exact.

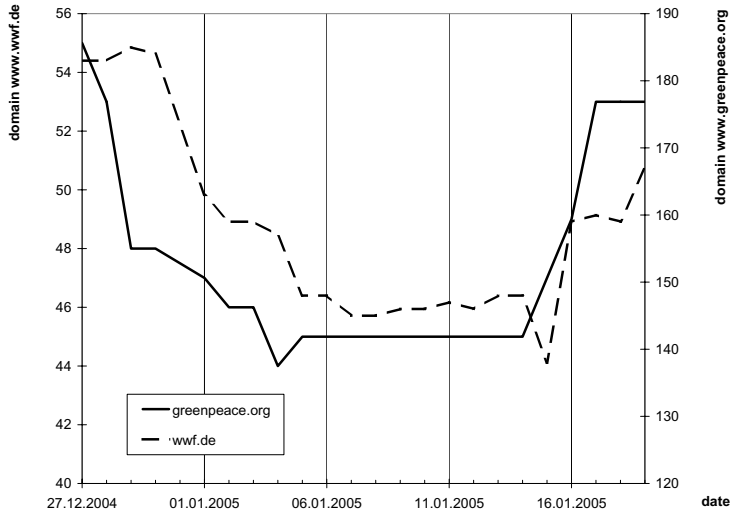

Fig. 2. Estimated hitcount values (Google) for 'Klimapolitik' (climate policy) on the domains `www.greenpeace.org` and `www.wwf.de` in time

Fig. 1 shows the developing of the visibility for the topics 'Weihnachten'[2] (Christmas) and 'Fasching' (Carnival) from Dec. 28, 2004 to Jan. 19, 2005. Obviously, the course of seasons leaves its traces on the internet. The visibility of 'Weihnachten' actually decreased by 25%. This is not a trivial finding for often web pages are created for a certain event but not necessarily removed afterwards, so we did not anticipate such a rapid decrease. The continuous growth of the web suggested that most of the webpages are kept.

The simultaneous change of visibility of one topic in different places is shown in Fig. 2, monitoring the topic 'Klimapolitik' (climate policy) from Dec. 27, 2004 to Jan. 19, 2005 in the two domains `www.greenpeace.org` and `www.wwf.de`. This clearly demonstrates the similarity of discussed topics among different sources. We will return to the example of climate policy below.

An impressive use case for the usability of the simple hitcount visibility measure in the context of marketing evaluation is described in [4] and should be mentioned at this point: in January 2005, a German company from the pharmaceutical branch (Dr.Kurt Wolff GmbH&Co. KG, brand name Alpecin) launched a new hair liquid called 'After Shampoo Liquid' with a special chemical compound as new ingredient, the 'Coffein-Complex'. There were marketing attempts in German media to promote this 'After Shampoo Liquid'. Commercials were emphasizing the 'Coffein-Complex' and encouraging consumers to visit the company's website and try the 'Glatzenrechner' ('balding calculator'[3]). A successful marketing campaign should draw public attention on the product and therefore

[2] All analysis for this paper was done in German.
[3] http://www.alpecin.de/en/balding-calculator/

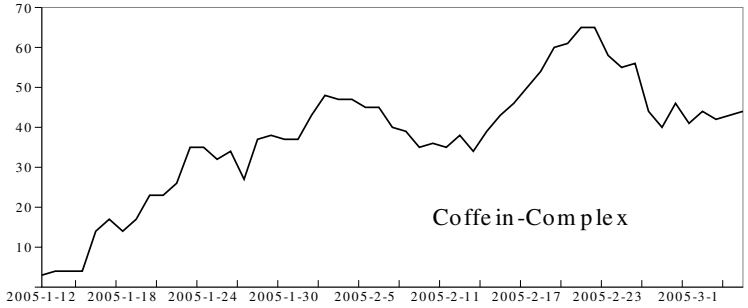

Fig. 3. Estimated hitcount values (Google) for 'Coffein-Complex' in time

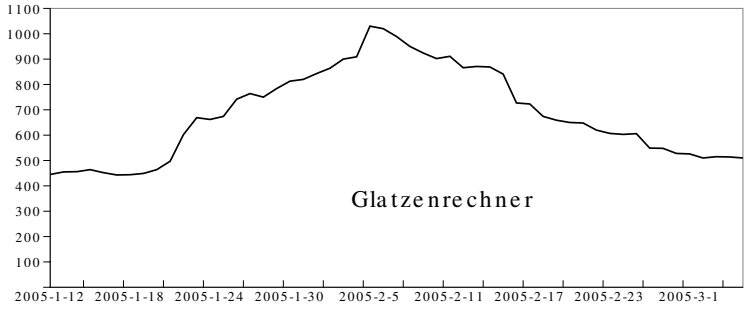

Fig. 4. Estimated hitcount values (Google) for 'Glatzenrechner' (balding calculator) in time

raise public visibility. We monitored the topics 'Coffein-Complex' (Fig. 3) and 'Glatzenrechner' (Fig. 4) from Jan. 12, 2005 to Mar. 5, 2005 and detected significant changes in visibility: 'Coffein-Complex' started with a hitcount of 3 and increased up to 65 on Feb. 22 before going down to the level of around 43. This shows how a product-related term or technology that did almost not exist on the internet can gain visibility through marketing actions. 'Glatzenrechner' was already present with a hitcount of 445, but more than doubled its hitcount to reach a maximum of 1030 on Feb. 5 before it approached a hitcount around 500.

Although the idea to measure visibility by hitcount values seems trivial and does not take the link structure or additional information into account, it has three main advantages:

1. It is based on existing search engines and therefore implemented quite easily.
2. It allows automated daily monitoring with only little effort.
3. It scales from monitoring visibilities from a certain domain to the whole (accessible) internet.

Defining topic visibility by the hitcount of one search term will hardly suit all use cases. Complex topics like 'US foreign policy during the cold war and its impacts on the German economy' often do not fulfill this requirement. However,

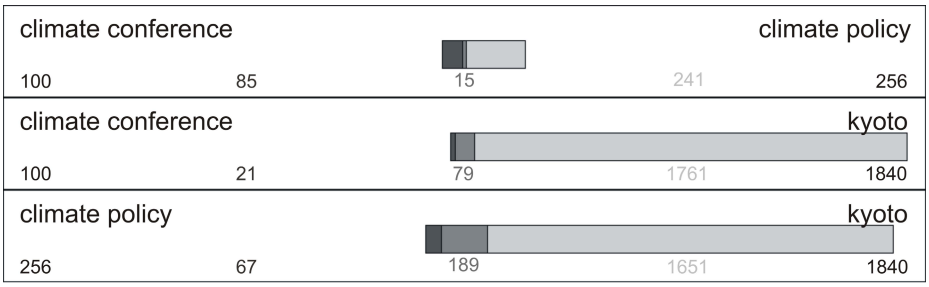

Fig. 5. Bar visualization of hitcount- and co-hitcount values (Google) for 'climate policy', 'climate conference' and 'Kyoto' (on `www.greenpeace.org`) at July 23, 2005. The graph shows the hitcount values (left/right), co-hitcount values (center) and the number of pages containing only one of two topics (left-centered/right-centered).

our analysis showed that it suffices for many cases and gives a useful base for the more complex models described in the following sections.

2.2 Co-visibility

To be able to describe dependencies between different topics we introduce the measure of co-visibility of two topics[4] based on co-occurrence: Two topics top_1 and top_2 co-occurring in a large number of documents should have *something* in common.[5] We measure the co-occurence with a co-hitcount value which we define as the hitcount of a search engine when searching for "top_1 AND top_2" (Fig. 5).

Again, an example from [4] illustrates how the co-hitcount of two topics can be used in a marketing scenario: in August 2005, all German carriers in the mobile phone market started offering flatrate contracts[6], called 'handyflatrate' (the German term for a mobile phone is 'handy'). Figure 6 shows the visibility of 'handyflatrate' from Aug. 3 to Aug. 24: it doubled in the beginning and returned to a hitcount of around 150.[7]

Figure 7 shows the co-hitcounts of the German main carriers T-Mobile, Vodafone, O2, E-Plus and Debitel with 'handyflatrate': all carriers gained visibility and it is obvious that the three biggest carriers generally had the highest values. However, the curve of E-Plus grows steadily and almost reaches that of the very big carrier Vodafone, while all other carriers settle on their level or even decrease. The same is revealed by Fig. 8 comparing the relative co-hitcounts ('attention shares') for carrier plus 'handy' and carrier plus 'handyflatrate' on Aug. 25. Note

[4] We restrict ourselves to two topics, generalization for three or more topics is possible.
[5] Whatever this "something" is. It is often *not* semantic closeness for authors not necessarily use synonyms within one text. So the interpretation of co-visibility has to be left to the user.
[6] 'Pay a constant amount of money per month and phone as long as you want'.
[7] All these monitorings were restricted on the domain 'de' to focus on the German market.

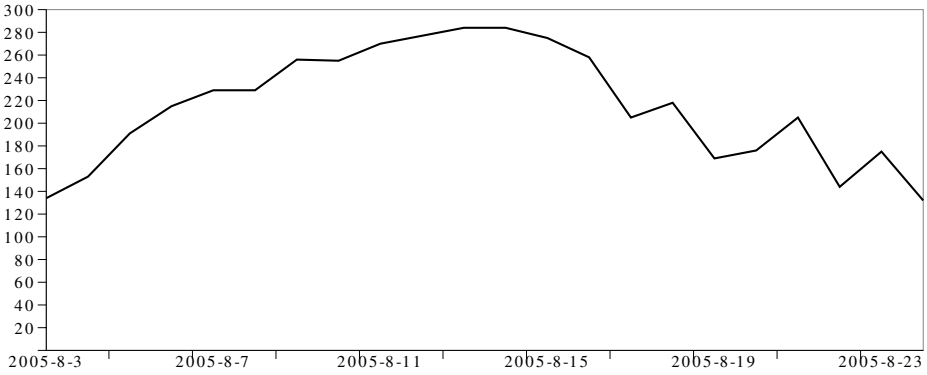

Fig. 6. Estimated hitcount (Google) for 'handyflatrate' on the domain 'de' in time

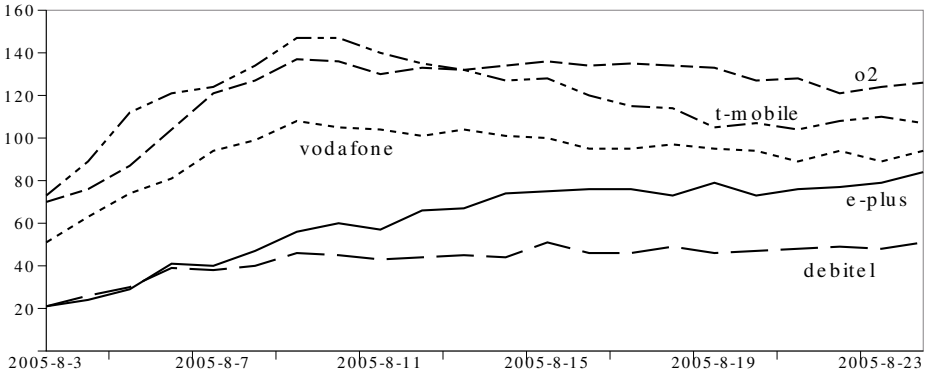

Fig. 7. Estimated co-hitcount (Google) for different carriers and 'handyflatrate' on the domain 'de' in time

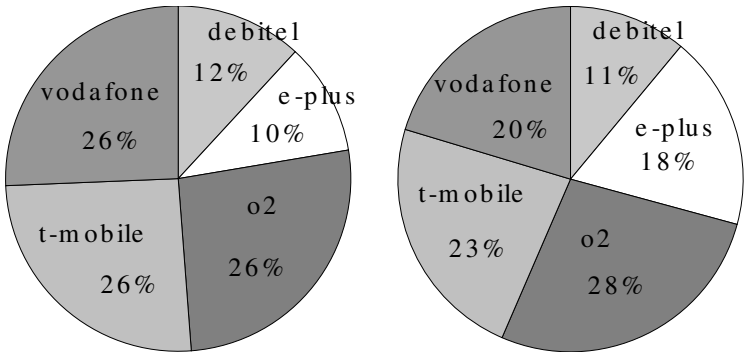

Fig. 8. Percentual co-visibilities of carriers and 'handy' (left), carriers and 'handyflatrate' (right) on the domain 'de' on Aug. 25

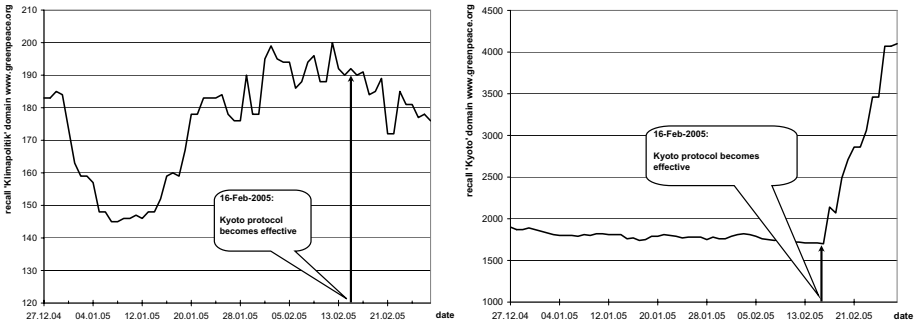

Fig. 9. Estimated hitcount values (Google) for 'Klimapolitik' (climate policy) [left] and 'Kyoto' [right] on `www.greenpeace.org`

that we could not use pure carrier hitcounts because of the special string 'O2' (we do not want to count pages related to oxygen). This diagram shows that E-Plus could increase their co-hitcount in the field of handyflatrates, compared to the overall hitcount of the company itself. This example illustrates how co-hitcounts can be used to link products with companies to analyze a market with different competitors under the aspect of public attention.

Anyhow, for many applications not the total number of pages is of interest, but the ratio between the number of pages containing both topics and the number of pages containing at least one of them. So we define

$$\text{covis}_i(top_1, top_2) \quad = \quad \frac{\text{cohitcount}(\text{“top}_1\text{”}, \text{“top}_2\text{”})}{\text{hitcount}(\text{“top}_i\text{”})}, \qquad i \in \{1, 2\}$$

which allows us to determine the degree of connection between several terms (currently or monitored in time).

3 Semantic Relations Between Topics

3.1 The Insufficiency of the Simple Visibility Measure

We tracked our example of climate policy in the domain `www.greenpeace.org` some further weeks and expected a rise in visibility on Feb. 16, 2005. At that date, 90 days after the ratification by Russia, the Kyoto protocol became effective. We expected important events like this to stimulate discussions on the topic climate policy and to be measurable in a domain dealing with environmental protection. Our results, pictured in Fig. 9 [left], did not support this hypothesis.

Contrariwise, the right side of Fig. 9 evidences an immense visibility gain for the topic 'Kyoto' in the same domain. This is easy to explain: an author writing an article for `www.greenpeace.org` reporting on the latest news on the Kyoto protocol will not necessarily use the phrase 'Klimapolitik', but definitely the word 'Kyoto'. On the other hand, doing without 'Klimapolitik' and monitoring only

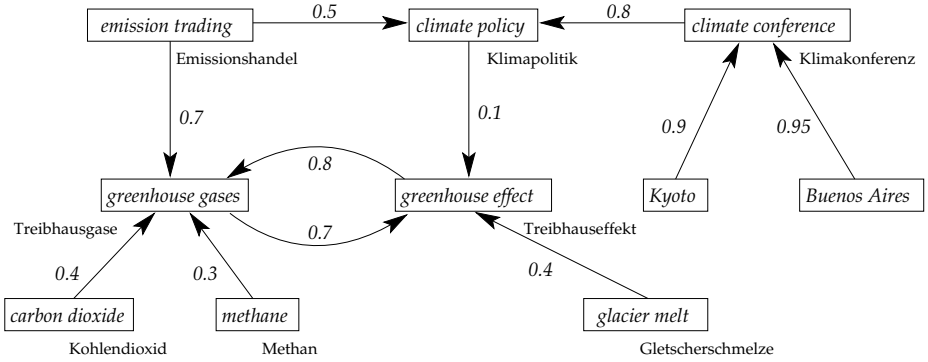

Fig. 10. Semantic network for topics of climate policy

'Kyoto' will not work likewise, for we cannot know a priori what *will* happen, so monitoring of at least the two topics 'Klimapolitik' and 'Kyoto' seems advisable. In general, this demonstrates the necessity to monitor more than one topic, more precisely several topics that are semantically related.

3.2 Semantic Network of Topics

We represent the following kind of relation between topics: two topics are semantically related, if the visibility of one topic automatically raises the visibility of the other. In other words: If a discussion on top_1 to a certain degree automatically concerns top_2, we designate top_1 as semantically related to top_2. Additionally, a weight $\mathrm{W}(top_1, top_2) \in [0, 1]$ qualifies the closeness of each relation with high values denoting a close relation. Take the topics HIV and aids as an example: A discussion on aids almost always also concerns HIV, for aids is always caused by the HI-virus. Actually, the two terms are quite often used synonymically. Further on, in the context of an environmental website, the topic Kyoto will rather reference the topic climate conference than the city of Kyoto, so a high semantical relation from Kyoto to climate conference exists. Note that our concept of semantical relationship is not symmetrical, e. g. a discussion on climate conference does not automatically as well concern Kyoto. Modeling the relations between several topics, we obtain a directed and weighted graph of topics, like illustrated in Fig. 10 for our example of climate policy. This graph corresponds to the well-known concept of semantic networks[8]. Keep in mind that the modeling of semantic topic networks certainly heavily depends on the context and the view of the modeler and cannot be specified objectively. In the case of the 0.9 between 'Kyoto' and 'climate conference' in Fig. 10, for example, this weight seems much too high for Kyoto might also refer to normal pages of the city Kyoto. But in the context of the domain www.greenpeace.org, Kyoto will almost always refer to a climate conference.

[8] See [5] for a comprehensive reading.

Although we regard visibilities as a general concept, the interpretation of visibilities as hitcount values like introduced in section 2 yet makes things clearer: An edge with weight $\mathrm{W}(\textit{Klimakonferenz}, \textit{Klimapolitik}) = 0.8$ claims that 80% of the web pages containing the string 'Klimakonferenz' as well concern the topic climate policy. Note that this is not a statement on co-visibility, i.e. those 80% may but need not necessarily contain the string 'Klimapolitik', but the results of a co-visibility request might help to build up the semantic network.

4 Spreading Activation with Co-visibilities

The algorithms we present in this section are based on the Spreading Activation algorithm (SA). SA was first introduced by psychologists as early as in the 1960's (see e. g. [6,7]) to explain human associative memory. Recently, SA was adopted for propagation of trust between actors in trust networks [8]. Furthermore, SA was utilized to improve methods in information retrieval (see e. g. [9,10,11]). The basic idea of SA is that of energy flowing through a network along weighted edges. Lausen and Ziegler specify the algorithm recursively (Alg. 1).

Algorithm 1 Spreading activation algorithm by Lausen and Ziegler [8].

procedure energize($e \in R_0^+, s \in V$) {

$\quad \mathrm{energy}(s) \longleftarrow \mathrm{energy}(s) + e$;

$\quad e' \longleftarrow \dfrac{e}{\sum_E \mathrm{W}(s,n)}$;

$\quad$ **if** $e > T$ **then** $\forall (s,n) \in E : \mathrm{energize}(e' \mathrm{W}(s,n), n)$;

}

V denotes the set of all nodes, E the set of all edges, s the node that is energized, e the amount of energy pushed into node s, $\mathrm{energy}(s)$ a data structure holding the current energy for each node (0 in the beginning), $\mathrm{W}(s,n)$ the weight of the edge from node s to node n. The energy a node s receives during one call of energize is disseminated proportionally on all outgoing edges of the node, depending on the accordant weight of the edge. This assures that not more energy than the injected energy e will leave the node. All nodes with incoming edges

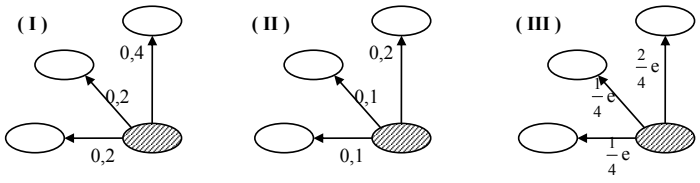

Fig. 11. Standard SA with energy spreading from the gray node

Algorithm 2 Spreading Activation algorithm for visibility adjustment

procedure visibilize($v \in R_0^+, t \in V$) {
  $\text{vis}(t) \leftarrow \text{vis}(t) + v$; ;
  if $v > T$ **then** $\forall (t,n) \in E : \text{visibilize}(v\,\text{W}(s,n), n)$;
}

from s are energized by a recursive call. Thus, energy packages with decreasing size flow through the network until their size falls under a certain threshold T and the algorithm terminates.

For the problem of visibility adjustment, a modification of this algorithm becomes necessary: Through the normalization of the outgoing energy, the graphs (I) and (II) in Fig. 11 become equivalent. This is contradictory to our intuition that a high semantic closeness between two topics should make more energy flow. Secondly, the assumption of SA that energy may not come from nothing, i. e. not more energy may leave a node than has been injected, is obsolete for visibilities. In fact, the notion of web pages concerning other pages implies some kind of 'hidden' visibility we strive to extract with our algorithm, so that a visibility gain is intended. We therefore simplify algorithm 1 and obtain algorithm 2, called visibilize for topic t and visibility v.

Algorithm 3 Spreading Activation algorithm with co-visibilities (1^{st} version)

procedure visibilize($v \in R_0^+, t \in V$) {
  $\text{vis}(t) \leftarrow \text{vis}(t) + v$; ;
  if $v > T$ **then** $\forall (t,n) \in E : \text{visibilize}(v\,\text{W}(s,n)(1 - \text{covis}_1(t,n)), n)$;
}

Algorithm 4 Spreading Activation algorithm with co-visibilities (2^{nd} version).

procedure visibilize($v \in R_0^+, t \in V, top_S \in V$) {
  $\text{vis}(t) \leftarrow \text{vis}(t) + v$; ;
  if $v > T$ **then** $\forall (t,n) \in E : \text{visibilize}(v\,\text{W}(s,n)(1 - \text{covis}_1(top_S, n)), n, top_S)$;
}

Using this algorithm, an adjustment of visibility is achieved as follows: model the semantic network of topics. Acquire the initial visibilities like described in section 2. For each topic t in the network call visibilize(t, v_{init}) with the initial visibility v_{init} of topic t, see Fig. 12 for an example with three topics and initial visibilities 100, 50, 10.

4.1 Spreading Activation with Co-visibilities

We do not settle for Algorithm 2, but improve it by adding knowledge from the co-visibilities. Imagine top_1 and top_2 from Fig. 12, with their initial visibilities of 100 and 50, having a $\text{covis}_1(top_1, top_2)$ of 0.4 and a $\text{covis}_1(top_2, top_1)$ of 0.8. In other words: 60 pages contain only the string of top_1, 10 pages only top_2, 40 pages contain both strings. Spreading the visibility of 50 from top_2 to top_1 and a visibility of 100 from top_1 to top_2 is not appropriate in this case, for some visibility would be counted double. We avoid this by introducing co-visibilities into our algorithm, refer to algorithm 3. Effectively, we adjust the weights of the net. Note that this adjustment is different for each date of monitoring, because the co-visibilities differ from day to day, while the original weights in the semantic network express the closeness of the relation and remain constant over time. Fig. 13 illustrates the first visibilization step for the new algorithm.

One aspect that is not covered by algorithm 3 is how to take cyclical and transitive relations into account for co-visibilities: In algorithm 3 we use for each package of energy propagated along an edge from top_1 to top_2 the co-visibility between top_1 and top_2. A more sophisticated strategy would take the co-visibility between the source topic top_S, i. e. the topic where the visibility has been injected, and top_2 (algorithm 4). The initial call of visibilize is executed with $t = top_S$. The recursive calls hand on the source-parameter and always use the co-visibility between source and the current target topic.

Going back to our example of climate policy, we run algorithm 4 on the initial visibility data of Fig. 9 with the semantic network of Fig. 10. We obtain adjusted visibilities for 'Klimapolitk', the topic we are interested in. Fig. 14 displays the

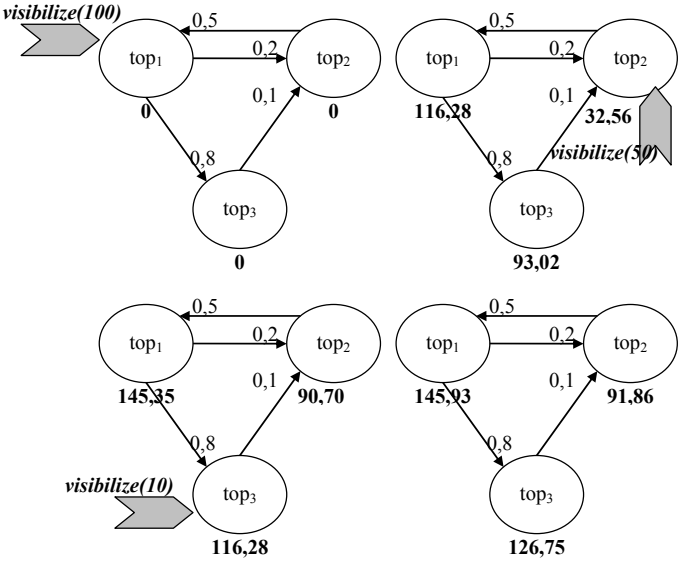

Fig. 12. Injection of visibility into a semantic network with three calls of algorithm 2

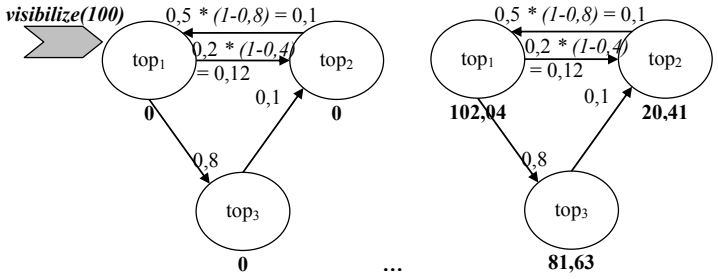

Fig. 13. Injection of visibility into a semantic network: first step with call of algorithm 3

initial visibilities of 'Klimapolitik' (lower curve), the initial visibilities of 'Kyoto' (center curve) and the adjusted visibilities of 'Klimapolitik' (upper curve). The developing of 'Klimapolitik' adapts itself to the developing of 'Kyoto'. This is no surprise, because we chose quite high weights in our semantic net leading to large packages of visibility flowing through the net.

5 Reliability of Search Engine Hitcounts

As mentioned in section 2, we obtained our hitcounts from the Google Web API, although the algorithms shown in this paper are not dependent on Google hitcounts, but also work with other sources of visibility. When we started our research on visibility in November 2004, the Google Web APIs seemed to be appropriate because of easy usability. However, Google state in their terms and

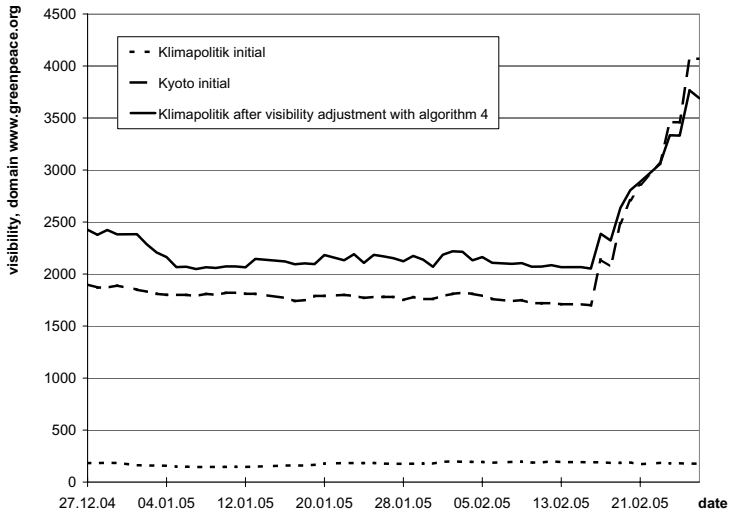

Fig. 14. Initial visibilities from Fig. 9 and adjusted visibilities for 'Klimapolitik' using the semantic network of Fig. 10 with Algorithm 4

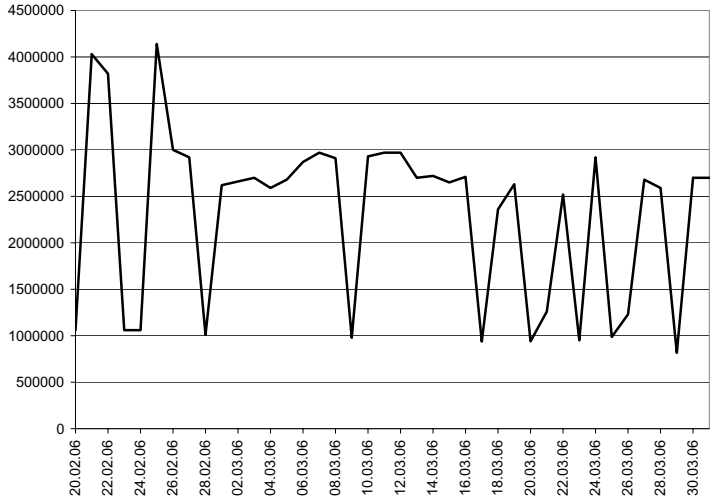

Fig. 15. Estimated hitcounts (Google API) of Dänemark (denmark)

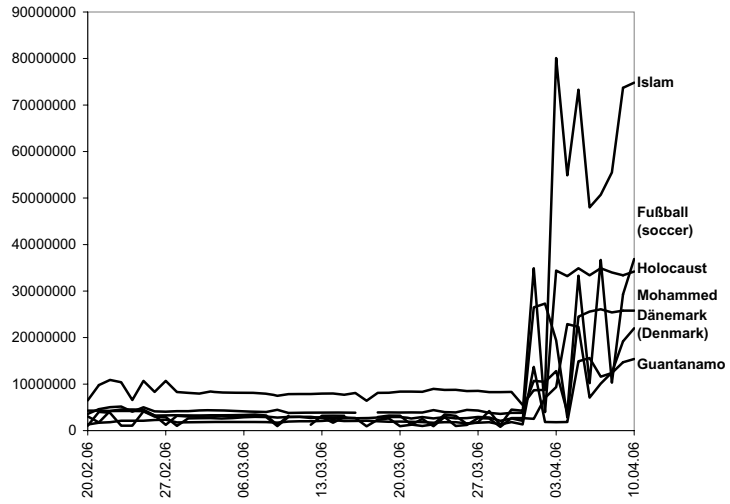

Fig. 16. Estimated hitcounts (Google API) of several topics increase tenfold on 1 April 2006

conditions for Google Web API service: 'The Google Web APIs service is currently in beta form and has not been fully tested or debugged' [9]. In general, using a search engine whose mechanisms you do not know in detail always implies relying on a black box. Other papers have reported on drawbacks of the

[9] `http://www.google.com/apis/api_terms.html`

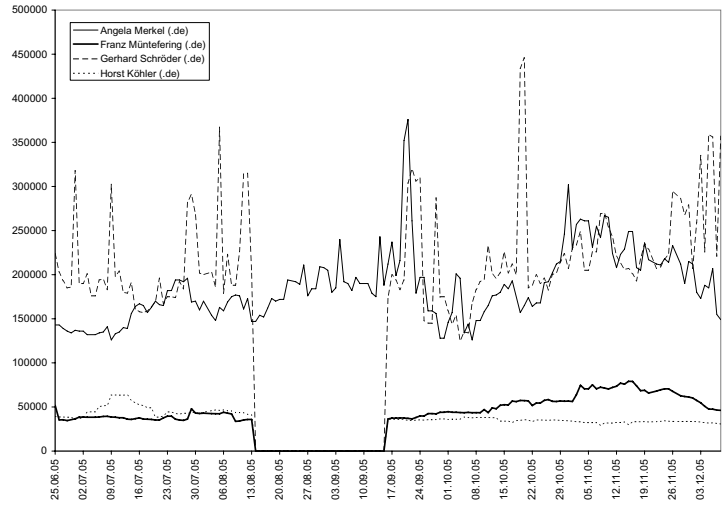

Fig. 17. Estimated hitcounts (Google API) of politicians with German umlauts

search engine's API, e.g. Mayr and Tosque who state that the hitcounts returned by the standard Google web interface and those from the Google API differ (see [12]). In his blog, Prof. Jean Véronis elaborates on some other inconsistencies of Google's hitcounts (see [13] and follow-up messages): for example, in January 2005 Google counted 8.000.000.000 pages containing 'the' on the whole internet, and 88.100.000 pages with 'the' on the English speaking web. That would imply that 99% of the pages containing 'the' were in other langugages than English.

In the course of our research, we experienced some further problems which come along with the Google Web API and which have a high impact on the API's reliability.

Consider Fig. 15 showing the hitcounts for 'Dänemark' (Denmark) in the course of one month: the values keep jumping from around 3.000.000 to 1.000.000 and back all the time. It is unrealistic to believe that the web has grown and shrunk so fast, so we can assume that some values are erroneous. Problems like this occured in some of our curves, especially in those with frequent topics like 'Denmark' or 'terrorism'. Although we knew a priori that we will not retrieve exact hitcounts, variations of this kind can hardly be seen as simply caused by estimation.

Figure 16 shows another problem which occured around Apr. 1, 2006: suddenly the hitcounts increased by factor ten (Fig. 16 only shows a small selection of the topics currently monitored). Finally, we had problems concerning German umlauts (ä, ö, ü) in August 2005: all topics with one of those letters dropped to almost zero for one month. This was particularly annoying because we monitored the hitcounts of German politicians like 'Schröder' (see Fig. 17) to evaluate the German elections in September 2005.

To cope with these problems hitcount data has to be handled carefully:

- A number of (independent) topics has to be monitored simultaneously to be able to detect hitcount changes caused by search engine internals (which should affect all topics).
- For many curves show dropouts (e. g. Fig. 15), single values are not reliable, so the average curve progression has to be used.
- Monitoring the hitcount of the same topics with more than one search engine can help to minimize the risk of data loss caused by errors like that of Fig. 17.
- No search engine is able to crawl the whole web and hitcount values are estimated, so the absolute numbers are rather erroneous. Therefore, the hitcount must not be interpreted absolutely. Instead, the relative change in time or the relation to the hitcounts of other topics should be used.
- The monitoring has to be supervised to detect failures like the umlaut problem, so the data should not be given to an uninformed end user.

6 Related Work and Conclusions

The first emphasis of our paper were the examples showing that real world events have an impact on the visibility of topics on the web. One problem yet remaining unresolved in this context is that, in contrary to our example of Kyoto, we often cannot predict which events could occur and which topics would be interesting to monitor. The terrorist attack of September 11, for example, surely had an effect on the visibility of 'terrorism' or 'World Trade Center', but nobody could know in advance that a monitoring of these topics would be interesting. The moment the event happens, the historical data is missing. A possible solution could be the usage of historical data from communication processes with timestamp, e.g. from a discussion group.

Analyzing the dynamically changing web has been done quite often: [14], for example, investigate the correlation between age of web pages and their quality to improve PageRank, while [15] monitor changes on the web to estimate the rate for reasonable search engine re-indexing. To our knowledge, no approach to correlate visibility and real world events exists.

In a second step, we modeled semantic relations between topics in semantic networks to add prior knowledge to our visibility analysis. Although these semantic networks look similar to Bayes Networks [16], Bayes networks do not permit the cyclic relations we need for modelling the semantic closeness of topics.

The approach of semantic networks was chosen to keep the algorithms simple. Nevertheless, an approach like thesauri with more than one type of relation would offer a much more intuitive modeling and therefore save time. In [17], a semi-automatical derivation of a semantic network from a user-modeled thesaurus is proposed. This would combine the intuitive modeling of thesauri with the convenience of a relatively simple algorithm for semantic networks.

The third input to our algorithm, besides visibilities and a semantic network, were co-visibilities. A possible application of co-visibilities we did not address in our paper is the automatic extraction of facts. This has recently been done

by Etzioni et al. who used hitcount values from a search engine for their system called KnowItAll to automatically extract facts from the WWW [18]. Search engine queries were also used by [19] for an automatical detection of synonyms and by [20] for the validation of question-answering systems, which both are further areas of application for co-visibilities. Co-occurences of terms are visualized in [21] for identifying significant topics in corpora of daily news.

We plan to endorse our findings on the relation between real world events and visibility with larger case studies. Investigations of at least more than one year should prove the applicability of visibility analysis in the long-term.

Finally, we intend to integrate the concept of visibility of topics into communication oriented modeling (COM) [22]. COM investigates large-scale communication processes with message/reference-networks like internet discussion groups. A definition of the concept of topic visibility for this kind of communication processes could be made. With the COM testing environment (COMTE)[10], further analysis could reveal correlations between author visibility, message visibility, topic visibility and the structure of the reference network.

References

1. Gans, H.J.: Deciding What's News: A Study of CBS Evening News, NBC Nightly News, 'Newsweek' and 'Time'. 25th anniversary edn. Northwestern University Press, Evanston, IL (2005)
2. Yungwook, K.: Measuring the economic value of public relations. Journal of Public Relations Research **13** (2001) 3–26
3. Falkinger, J.: Attention economies. CESIFO WORKING PAPER NO. 1079, ifo Institut für Wirtschaftsforschung, München (2003)
4. Kiefer, P., Stein, K.: Visibility analysis on the web as an indicator for public relations and marketing evaluation. In: Proc. of Intelligent Agents, Web Technology and Internet Commerce (IAWTIC 2005), IEEE Computer Society Publishing (2005)
5. Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA (2000)
6. Quillian, R.: Semantic memory. In Minsky, M., ed.: Semantic Information Processing. MIT Press, Boston, CA, USA (1968) 227–270
7. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review **82** (1975) 407–428
8. Lausen, G., Ziegler, C.N.: Spreading activation models for trust propagation. In: IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE '04. (2004) 83–97
9. Preece, S.: A spreading activation network model for information retrieval. PhD thesis, CS Dept., Univ. of Illinios, Urbana, IL. (1981)
10. Crestani, F.: Applications of spreading activation techniques in information retrieval. Artificial Intelligence Review **11** (1997) 453–482
11. Ceglowski, M., Coburn, A., Cuadrado, J.: Semantic Search of Unstructured Data using Contextual Network Graphs. National Institute for Technology and Liberal Education (2003)

[10] `http://www.kinf.wiai.uni-bamberg.de/COM/`

12. Mayr, P., Tosques, F.: Google Web APIs - an instrument for webometric analyses? Poster presented at ISSI 2005 (2005) http://arxiv.org/ftp/cs/papers/0601/0601103.pdf.
13. Véronis, J.: Web: Google's counts faked? Blog, see also follow-up messages (2005) http://aixtal.blogspot.com/2005/01/web-googles-counts-faked.html.
14. Baeza-Yates, R., Saint-Jean, F., Castillo, C.: Web structure, age and page quality. In: Proceedings of the 2nd International Workshop on Web Dynamics (WebDyn 2002). (2002) http://www.dcs.bbk.ac.uk/webDyn2/onlineProceedings.html.
15. Brewington, B.E., Cybenko, G.: How dynamic is the Web? Computer Networks (Amsterdam, Netherlands: 1999) **33** (2000) 257–276
16. Russel, S., Norvig, P.: Chapter 14, Probabilistic Reasoning. In: Artificial Intelligence: A Modern Approach. Prentice Hall (2003) 492–536
17. Kiefer, P.: Computational analysis of the visibility of themes in internet-based communication processes, in German: Softwaregestützte Analyse der Sichtbarkeit von Themen in internetbasierten Kommunikationsprozessen. Diploma thesis, Chair for Computing in the Cultural Sciences, Bamberg University, Bamberg (2005)
18. Etzioni, O., Cafarella, D., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence (2005) 91–134
19. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of ECML2001, Freiburg, Germany (2001) 491–502
20. Magnini, B., Negri, M., Tanev, H.: Is it the right answer? Exploiting web redundancy for answer validation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. (2002) 425–432
21. Richter, M.: Analysis and visualization for daily newspaper corpora. In: Proc. of RANLP. (2005) 424–428
22. Malsch, T., Schlieder, C., Kiefer, P., Lübcke, M., Perschke, R., Schmitt, M., Stein, K.: Communication between process and structure: Modelling and simulating message-reference-networks with COM/TE. Journal of Artificial Societies and Social Simulation (2006) accepted.