# Visibility Analysis on the Web Using Covisibilities 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. We 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 recall 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 covisibility between topics is integrated to obtain an algorithm that mostly complies with an intuitive view on visibilities.

## 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, 3], 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. [4]). 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 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 recall values of a search engine, present examples indicating that real world events actually do have an impact on visibility on the web and introduce the concept of topic covisibility (section 2).

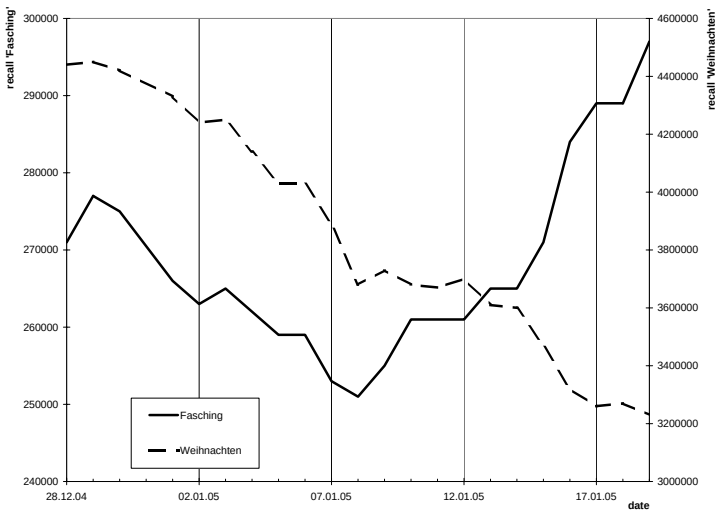


**Fig. 1.** Estimated recall values (Google) for 'Weihnachten' (christmas) and 'Fasching' (carnival) in time.

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).

At last (section 5) we give a short summary and a view on current research.

## 2 Visibility and Covisibility

### 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 $\text{vis}(top) = \text{recall}(\text{“top”})$ with $\text{recall}(\text{“top”})$ being the number of pages found on the search term "top" by a given search engine.[1].

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

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

[2] All analysis for this paper was done in German.

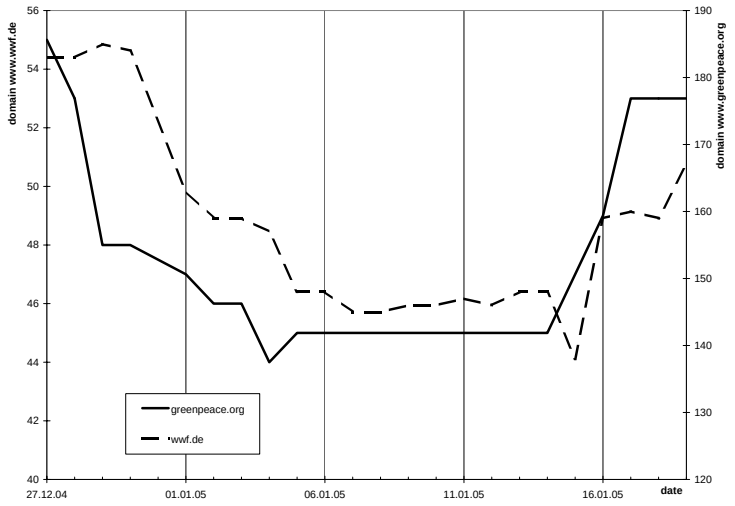


**Fig. 2.** Estimated recall values (Google) for 'Klimapolitik' (climate policy) on the domains www.greenpeace.org and www.wwf.de in time.

'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 (especially during christmas vacation), 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 '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 our example of climate policy below.

Although the idea to measure visibility by recall 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 recall 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, our analysis showed that it suffices for many cases and gives a useful base for the more complex models described in the following sections.

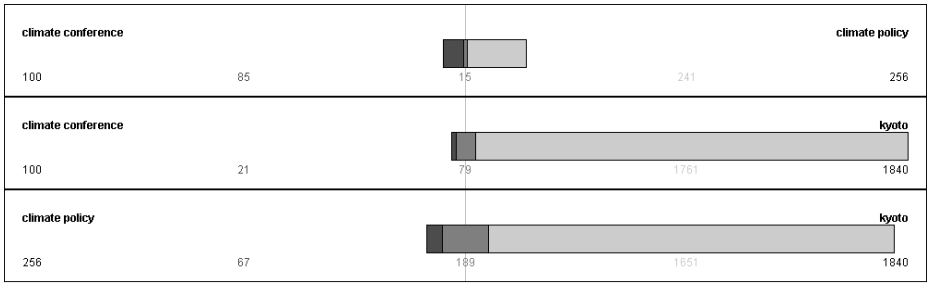


**Fig. 3.** Bar visualization of recall- and co-recall values (Google) for 'climate policy', 'climate conference' and 'Kyoto' (on www.greenpeace.org) at July 23, 2005. The graph shows the recall values (left/right), co-recall values (center) and the number of pages containing only one of two topics (left-centered/right-centered).

### 2.2 Covisibility

To be able to describe dependencies between different topics we introduce the measure of covisibility of two topics[3] based on co-occurrence: Two topics $top_1$ and $top_2$ cooccurring in a large number of documents should have *something* in common.[4] We measure the co-occurence with a co-recall value which we define accordingly to recall as the hit count of a search engine when searching for "$\text{top}_1$ AND $\text{top}_2$" (Fig. 3).

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

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

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. 4 [left], did not support this hypothesis.

[3] we restrict ourselves to two topics, generalization for three or more topics is possible.

[4] whatever this "something" is. It is often *not* semantic closeness for authors not necessarily use synonyms within one text. So the interpretation of covisibility has to be left to the user.

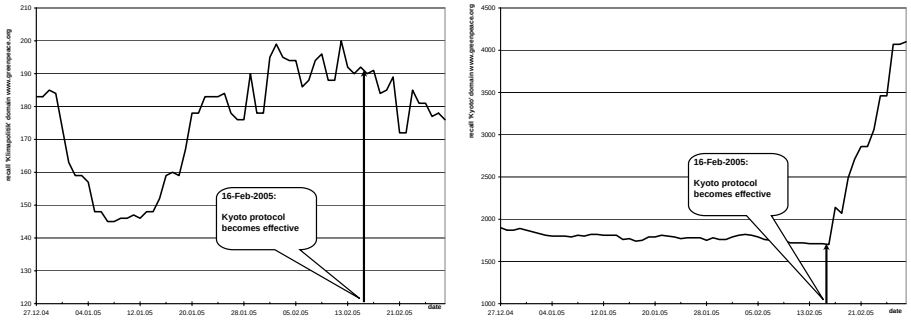


**Fig. 4.** Estimated recall values (Google) for 'Klimapolitik' (climate policy) [left] and 'Kyoto' [right] on www.greenpeace.org.

Contrariwise, the right side of Fig. 4 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 '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. 5 for our example of climate policy. This graph corresponds to the well-known concept of semantic networks[5]. Keep in mind that the modeling of semantic topic networks certainly heavily depends on the

[5] see [5] for a comprehensive reading

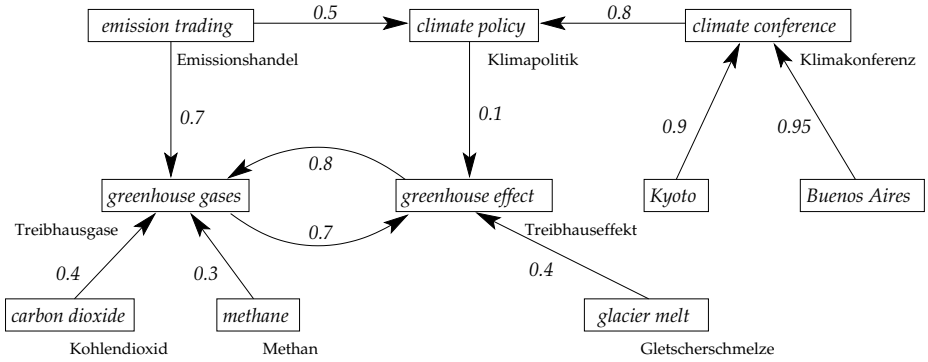


**Fig. 5.** Semantic network for topics of climate policy.

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. 5, 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.

Although we regard visibilities as a general concept, the interpretation of visibilities as recall 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 covisibility, i.e. those 80% may but need not necessarily contain the string 'Klimapolitik', but the results of a covisibility request might help to build up the semantic network.

## 4 Spreading Activation with Covisibilities

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. [9–11]). 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).

$V$ 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 $s$, 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 from $s$ are energized by a recursive call. Thus, energy packages with decreasing

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

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

energy($s$) $\longleftarrow$ energy($s$) + $e$ ;

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

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

}

**Algorithm 2** Spreading Activation algorithm for visibility adjustment

**procedure** visibilize($v \in R_0^+, t \in V$) {

vis($t$) $\leftarrow$ vis($t$) + $v$; ;

**if** $v > T$ **then** $\forall (t,n) \in E$ : visibilize($v\,\mathrm{W}(s,n), n$) ;

}

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. 6 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 $v$.

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($t, v_{\mathrm{init}}$) with the initial visibility $v_{\mathrm{init}}$ of topic $t$, see Fig. 7 for an example with three topics and initial visibilities 100, 50, 10.

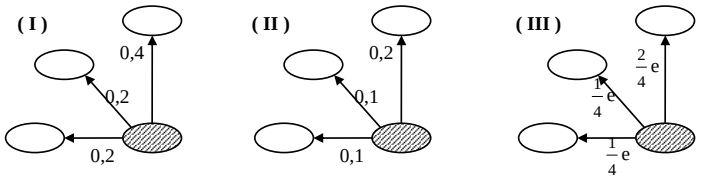


**Fig. 6.** Standard SA with energy spreading from the gray node.

**Algorithm 3** Spreading Activation algorithm with covisibilities ($1^{\text{st}}$ 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 covisibilities ($2^{\text{nd}}$ 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** $\forall (t,n) \in E : \text{visibilize}(v\, \text{W}(s,n)(1 - \text{covis}_1(top_S,n)), n, top_S)$ ;

}

### 4.1 Spreading Activation with Covisibilities

We do not settle for Algorithm 2, but improve it by adding knowledge from the covisibilities. Imagine $top_1$ and $top_2$ from Fig. 7, 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 covisibilities 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 covisibilities 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. 8 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 covisibilities: In algorithm 3 we use for each package of energy propagated along an edge from $top_1$ to $top_2$ the covisibility between $top_1$ and $top_2$. A more sophisticated strategy would take the covisibility 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 covisibility 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. 4 with the semantic network of Fig. 5. We obtain adjusted visibilities for 'Klimapolitk', the topic we are interested in. Fig. 9 displays the 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

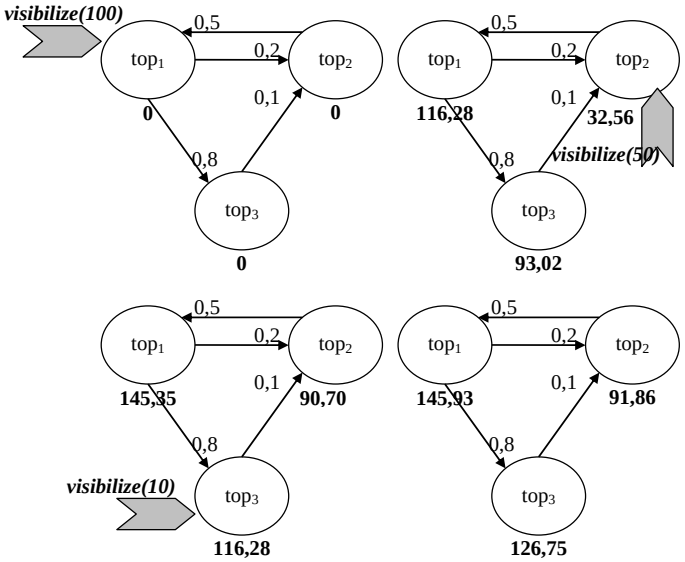


**Fig. 7.** Injection of visibility into a semantic network with three calls of algorithm 2.

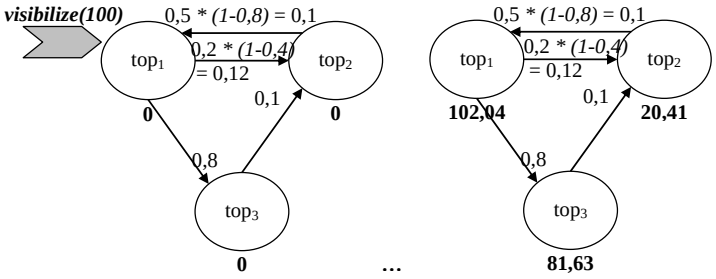


**Fig. 8.** Injection of visibility into a semantic network: first step with call of algorithm 3.

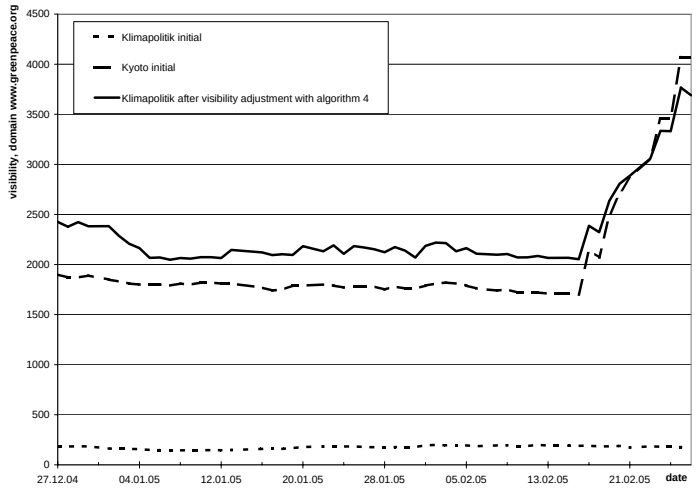


**Fig. 9.** Initial visibilities from Fig. 4 and adjusted visibilities for 'Klimapolitik' using the semantic network of Fig. 5 with Algorithm 4.

no surprise, because we chose quite high weights in our semantic net leading to large packages of visibility flowing through the net.

## 5 Conclusion and Outlook

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 data from communication processes with timestamp, e.g. from a discussion group, and to somehow extrapolate this historical data with the current visibilities monitored after the event.

Analyzing the dynamically changing web has been done quite often: [12], for example, investigate the correlation between age of web pages and their quality to improve PageRank, while [13] 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. 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 [14], 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 semantic network, were covisibilities. A possible application of covisibilities we did not address in our paper is the automatic extraction of facts. This has recently been done by Etzioni et al. who used recall values from a search engine for their system called KnowItAll to automatically extract facts from the WWW [15–17]. Furthermore, they used recall values to improve the precision of an information extraction system [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 covisibilities. Anyhow, what none of these approaches addressed is the monitoring of visibilities over time with respect to real-world events.

We plan to endorse our findings on the relation between real world events and visibility with larger case studies. Currently we are monitoring parties, politicians and political topics for the German election in September 2005. We hope to be able to retrace daily events of the election campaign with visibilities. Possibly there might also be a match between the developing of visibilities for political topics/politicians and findings from election researchers (e.g. why did people elect a certain party). With another case study we will cover the area of marketing/public relations to get an indicator for the evaluation of public relation actions. A cooperation with scientists from these areas for evaluation purposes is prefered. Furthermore, 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) [21]. 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)[6], 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. Hon, L.C.: Demonstrating effectiveness in public relations: Goals, objectives, and evaluation. Journal of Public Relations Research **10** (1998) 103–135
3. Yungwook, K.: Measuring the economic value of public relations. Journal of Public Relations Research **13** (2001) 3–26


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

4. Falkinger, J.: Attention economies. CESIFO WORKING PAPER NO. 1079, ifo Institut für Wirtschaftsforschung, München (2003)
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)
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)
12. 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)
13. Brewington, B.E., Cybenko, G.: How dynamic is the Web? Computer Networks (Amsterdam, Netherlands: 1999) **33** (2000) 257–276
14. 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)
15. Etzioni, O., Cafarella, D., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., Yates, A.: Web-scale information extraction in knowitall (preliminary results). In: Proceedings of the 13th Intl. World Wide Web Conference. (2004)
16. Etzioni, O., Cafarella, D., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D., Yates, A.: Methods for domain-independent information extraction from the web: an experimental comparison. In: Proceedings of the 19th National Conference on Artificial Intelligence. (2004)
17. 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)
18. Soderland, S., Etzioni, O., Shaked, T., Weld, D.: The use of web-based statistics to validate information extraction. In: Papers from the AAAI-2004 Workshop on Adaptive Text Extraction and Mining(ATEM-2004), San Jose, CA (2004)
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. Malsch, T., Schlieder, C.: Communication without agents? from agent-oriented to communication-oriented modeling. In: Regulated Agent-Based Social Systems: First International Workshop, RASTA 2002, Bologna, Italy, July 16, Springer-Verlag, Berlin, Heidelberg, New York (2004) 113–133