HAL
open science

Shared conceptualisations in weblogs

Anjo Anjewierden, Rogier Brussee, Lilia Efimova

To cite this version:

Anjo Anjewierden, Rogier Brussee, Lilia Efimova. Shared conceptualisations in weblogs. BlogTalks 2.0: The European Conference on Weblogs, 2004, Vienna, Austria. pp.110-138. hal-00190692

HAL Id: hal-00190692

https://telearn.hal.science/hal-00190692

Submitted on 23 Nov 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Shared Conceptualisations in Weblogs

Anjo Anjewierden, anjo@swi.psy.uva.nl, http://anjo.blogs.com
University of Amsterdam, Social Science Informatics, The Netherlands

Rogier Brussee, rogier.brussee@telin.nl, http://rogierbrussee.blogspot.com.
Lilia Efimova, lilia.efimova@telin.nl, http://blog.mathemagenic.com
Telematica Instituut, Enschede, The Netherlands

Abstract. In this paper we investigate how conceptualisations can be identified in weblogs using language technology (automated text analysis). We focus on getting a handle on both the concepts bloggers use and the way they think these concepts are related. The analysis of these conceptualisations can then be applied to a single weblog, resulting in a visualisation of potential conceptualisations the blogger wants to share with the outside world. Another type of analysis is to determine the overlap, or sharedness, of conceptualisations between bloggers. We have implemented both analysis approaches in an interactive tool.

1. Introduction

Several studies have analysed blogging as a social activity by studying blogger characteristics (gender, age, background), through characteristics of posts (age of the blog, frequency and length of posts) and by the interconnectedness of blogs (for example, Aïmeur, Brassard and Paquet, 2003; Herring, Scheidt, Bonus and Wright, 2004; Nardi, Schiano and Gumbrecht, 2004; Nilsson, 2003). In this paper we take a further step and study blogs as a platform to share conceptualisations.

Blogs are seen as a means to share information, opinions and knowledge. To do this, bloggers partially share terminology and refer to each other's posts. However, it is not immediately clear that the use of a shared term also implies that conceptualisations are shared. In other words it is not clear that bloggers have the same mental image of the meaning of the shared terms. Such misconceptions might easily arise as bloggers of totally different backgrounds can meet each other in the blogosphere.

Shared conceptualisations are important for communication and learning because these conceptualisations are the basis for understanding and dialogue. We believe weblogs are a unique vehicle for defining and exposing personal conceptualisations because the time between publishing and discussion seems to be much smaller and the diversity of people exposed to these ideas seems to be much larger than for traditional publishing and project teams.

In this paper we investigate how conceptualisations can be identified in weblogs using language technology (automated text analysis). We focus on getting a handle on both the concepts bloggers use and the way they think these concepts are related. The analysis of these conceptualisations can then be applied to a single weblog, resulting in a visualisation of potential conceptualisations the blogger wants to share with the outside world. Another type

of analysis is to determine the overlap, or sharedness, of conceptualisations between bloggers. We have implemented both analysis approaches in an interactive tool.

Section 2 describes the overall approach of our work. Section 3 provides details about the techniques and algorithms used to obtain the data to be analysed. Section 4 describes the functionality of a tool, called *Sigmund*, which allows interactive browsing of the conceptualisations found. Section 5 gives some examples based on the analysis of blogs that are mainly about Knowledge Management. Section 6 discusses our work in relation to other research. Finally, Section 7 contains the conclusions.

2. Approach

Over the last decades there has been a significant amount of research into capturing conceptualisations with an emphasis on using formal and machine inspectable representations. A primary result of this research is the notion of ontologies (Gruber,1993; Staab and Studer, 2004), which has resulted in knowledge representation languages such as RDFS and OWL and the idea of a Semantic Web.

An interesting research issue is whether it is possible to take some body of text (e.g. an article, textbook or blog) and automatically extract the conceptualisations it contains. In general this problem has proven much too hard as it not only requires reading and parsing the text, but also understanding the meaning and having the subtle sense of social context that allows to distinguish serious attempts to structure knowledge from opinions, lucid observations, humor and hobbies. This is even more difficult than machine translation, as it involves a model of the physical and social world that will be difficult to obtain without actual involvement in that world. Moreover, our aim is to discover the relations between concepts bloggers seem to perceive themselves, rather than imposing our own model of the world. The approach therefore settles on using "observations" of blogger's use of terms and relatively crude statistical analysis.

The main difference between our research and the work on ontologies and the Semantic Web is that the latter tries to formalise structures that are believed to exist in some abstract sense in the real world, thereby making classification and inferencing possible (e.g. that apples are fruit). On the other hand, we have developed techniques that pick up the patterns people leave in their weblog which we believe are a result of their use of an underlying conceptualisation. The resulting representations of these conceptualisations may or may not be shared or "true". In summary, the approach is as follows:

1. Identify terms that potentially point to concepts. Here we make a distinction between terms that point to names of people and terms that represent the subject matter of the text. See also Section 3.2.

2. Once the concepts have been identified, we need to establish whether they are semantically related, at least according to the blogger. For this we rely on the assumption that there is some sort of semantic relation between terms if these terms are often used together in the same post. As an operationalisation of the strength of the relation we use a statistical measure for "the risk" of using one term given that a blogger is using another term in the same post together with an estimate for the chance that a pattern is merely a coincidence. See Section 3.3.

3. The output of the above is a two dimensional table of “risks” for using terms in combination. A typical weblog contains thousands of terms, so the table contains millions of entries. We then select the high “risk” combinations and graphically represent the resulting clusters of terms for the end-user. Preliminary experience shows that visual inspection and human experience often suggest an underlying semantic relation between terms. See Section 3 for the methodology and Section 5 for examples.

3. Methodology

This section describes the language related methodology underlying the approach to collecting weblog data. In particular, we describe the web spidering method (Section 2.1), the language technology used to find interesting terms in weblogs (Section 2.2) and the algorithms to calculate similarity between terms (Section 2.3).

3.1 Spidering weblogs

The base data required is a significant fraction of complete posts of a weblog. Unfortunately RSS feeds cannot be used for this. First, RSS feeds typically capture no more than the latest 15 posts or so, and secondly most RSS feeds do not contain complete posts. There also does not appear to be a tool that performs the task of extracting all posts from a weblog. The solution we have settled on, is to spider weblogs using the available HTML pages on the web and subsequently use heuristics to extract the posts from these pages. Thereafter, the extracted posts are converted to the RSS 2.0 format and fed into the language technology module (Section 3.2).

A brief description of the heuristics used for spidering is given below. The process is initiated by the user providing the URL of a weblog to be analysed. There are two basic steps:

1. Locate links to the archives of the weblog; and
2. Extracting the posts from these archive pages.

The first step turns out to be relatively easy as most blogging software names archives consistently using the year, month and optionally the day. The second step is more complicated and error prone as blogging software allows significant control over the presentation of a weblog post as an HTML page. From the definition of a weblog post in RSS it must at least contain: date of publication, title, body (i.e. the post itself) and a permalink. Patterns to detect these have been incrementally developed and we empirically believe more than 90% of (English) weblog archives in HTML are correctly converted to RSS by our blog spider.

3.2 Language Technology

Compared to more carefully edited publications such as articles in papers, magazines or journals, weblog posts can be characterised as "noisy". Noise in posts consists mainly of misspellings, alternate spellings that are not common (e.g. *weblog* vs. *blog*, or *on-line* vs. *online*) and an abundance of abbreviations (e.g. *KM* for *Knowledge Management*).

Standard language technology can correct most misspellings if the edit distance is one (i.e. if two characters are transposed, , one character is inserted or deleted, or diacritics are not used). Alternate spellings are properly handled if they appear in the dictionary. Synonyms, for example weblog and blog, have to be provided by the user.

The two main challenges with respect to language technology are the identification of meaningful terms and the extraction of the names of people, organisations and things. The latter is part of a well-known Information Extraction task called the recognition of "named entities" and is addressed in the literature.

We define a **meaningful term** to be a (possibly compound) term that refers to a single concept irrespective of the specific textual rendering. The **(semantic) term class** of a meaningful term is defined as the set (equivalence class) of all terms referring to the same concept and is denoted with square brackets around the term. The first task is therefore to find meaningful terms and the second problem is to collect the terms that belong to the same term class. Often a meaningful term corresponds linguistically to a noun phrase: a term constructed from a sequence of consecutive nouns (*weblog post, knowledge management*). Because of the definition of a term class, the meaningful terms *KM* and *knowledge management* are both members of the same term class [*knowledge management*] as they refer to the same concept (provided of course that KM is an abbreviation for knowledge management in a given weblog). Similarly, inflected forms (e.g. plurals, past tense), misspellings, alternate spellings and user provided synonyms are also treated as members of a term class.

The analysis of a weblog proceeds as follows:

1. Identify potential terms. The algorithm scans over the posts and collects all sequences of words separated by stop words. For example, the sentence: *This is Knowledge Management Research ...*, results in the following potential meaningful terms being recorded: *knowledge, management, research, knowledge management, management research* and *knowledge management research*. These terms are then normalised using the CELEX dictionary (Baayen et al., 1995), for example *supporting informal learning* becomes *support informal learn*.

2. Expand abbreviations. The second step in processing a weblog is expanding the short forms of abbreviations to their corresponding long forms. Because of the noisy nature of weblogs traditional abbreviation finding algorithms (e.g. Schwartz & Hearst, 2003) that rely on the short and long forms appearing next to each other do not work. The algorithm we use is based on the idea that the long form must be a meaningful term and that both the long and the short forms appear relatively frequently. A stop list of very common abbreviations (e.g. PC, CD, OS, etc.) is used to prevent accidental expansions. The outcome is that all occurrences of the short form of an abbreviation are treated as if the long form appeared. This procedure is also applied recursively. For example, *KM summer school* and *KMSS* are all treated as linguistic variants of term class [*knowledge management summer school*].

3. Normalise terms. Our definition of a meaningful term excludes prefixing terms with adjectives that refer to specific points in space and time, or qualitative and quantitative observations. For example, the term *today's post* is reduced to *post*, and *good blog* is reduced to *blog*. This step is implemented by using a list of words to be excluded as prefixes.

4. Delete implied and low frequency terms. The next step is to delete all terms that are implied by longer terms. For example, if all occurrences of *management research* are part of *knowledge management research* then the former is redundant and can thus be

ignored. A term has to appear at least four times in a given weblog to be considered for analysis.

More text analysis techniques can obviously be added. An extension we are considering is treating conflations (e.g. *knowledge management* vs. *management of knowledge*) as synonyms as they appear heavily used in weblogs.

3.3 Co-occurrence

Intuitively we will say that term B **co-occurs** with term A if the frequency of term B in posts containing term A is much higher than the frequency of term B in posts not containing term A. Even on this intuitive level it is clear that co-occurrence is not symmetric. For example, we could find that for a given weblog, *knowledge* co-occurs with *management*. This would happen if a blogger uses the term *knowledge* (almost) always in the combination *knowledge management*. However, this blogger may well blog often about *management* in general and use the term *knowledge management* very seldom. Therefore the frequency of the term *management* need not be much elevated if the term knowledge occurs, in fact it conceivably might be reduced.

Clearly "much higher" is an insufficiently precise notion and we need to make quantitative statements. We therefore use the following model. Given a term A we separate blog posts in two groups: those containing A and those not containing A. In each of the groups we count the number of occurrences of the term B. Since the A group and the not A group may be of very different sizes, we normalise each number of occurrences of B by the number of terms the groups contain. This leads to the following:

Definition: Let $n(B \mid A)$ (respectively $n(B \mid \neg A)$) be the number of occurrences of the term B in posts that contain the term A (respectively do not contain the term A), and likewise let $n(* \mid A)$ (respectively $n(* \mid \neg A)$) be the total number of terms in the posts that contain the term A (respectively do not contain the term A)). Then the **co-occurrence degree** $c(B \mid A)$ is defined as the number

$$c(B \mid A) = \frac{n(B \mid A) / n(* \mid A)}{n(B \mid \neg A) / n(* \mid \neg A)}$$

We say that B co-occurs with A to degree k if $c(B \mid A) \geq k$.

Note that $c(B \mid A) = 1$ if B is as frequent in posts containing A as it is in posts not containing A i.e. that term B and A seem to be unrelated. Also note that $c(B \mid A) < 1$ means that the use of term A tends to discourage the use of B. Sigmund tool allows the setting of the co-occurrence degree. A reasonable default is a factor of 4.0.

All counting methods and derivatives like co-occurrence suffer from "statistical uncertainty". While it is possible to claim beyond reasonable doubt that terms occur a given number of times in a selected sample of posts, we should not attach any significance to two terms completely co-occurring just because these terms occur once in a single post. After all we have only selected a sample of posts and the posts only represent a sample of the sentences that those blogger(s) ever wrote, let alone spoke. Therefore, it is important to have an estimate for the reliability of our co-occurrence measure. Such estimates cannot be made without making assumptions about the nature of blog posts as "random samples".

We will assume a probabilistic model of posts as random streams of terms that with a certain probability contain term B. What we want to determine is how much larger the probability to contain B is for posts containing A than it is for those not containing A.

This model is similar to the one used for the relative risk statistics (Daniel 1995, pp. 542-555; Sabo, 2003). The normal use of the relative risk is to estimate how much more likely a smoker is getting cancer than a non-smoker. Co-occurrence can be seen as an estimate of how much more a blogger is "at risk" of using term B given that s/he indulges in the "risky behaviour" of using term A. The statistics community has analysed this situation and comes up with an estimate for the confidence interval of the co-occurrence. Assuming that the numbers $n(B \mid A)$, $n(B \mid \neg A)$ (and $n(* \mid A)$, $n(* \mid \neg A)$) are not too small, at least on the order of 5 (for $\alpha = 5\%$ range to use χ^2 statistics) at the $100(1-\alpha)\%$ confidence level the co-occurrence level is *at least*

$$c(B \mid A)^{1-z(\alpha)/\sqrt{X^2}}$$

Here $z(\alpha)$ is the z-value of α (i.e. the value of z such that the area under the normal distribution *above* z is α). For example, for $z(5\%) = 1.6452$. The value of X^2 is then estimated as

$$X^2 = n\frac{(n(B \mid A)n(\neg B \mid \neg A) - n(B \mid \neg A)(n(\neg B \mid A))^2}{n(B)n(\neg B)n(A)n(\neg A)}$$

Where n is the total number of selected terms in all the posts, and $n(A)$ (respectively $n(\neg A)$) is the number of occurrences of A (respectively of selected terms other than A). Note that X^2 scales roughly like n, so the exponent $1 - z(\alpha)/\sqrt{X^2}$ tends to 1 roughly like $1/\sqrt{n}$.

4. Tool Support

The approach and methodology outlined in the previous sections have been implemented in a tool called *Sigmund*, after Sigmund Freud the psychiatrist from Vienna. The name is inspired by the observation that bloggers who use the tool on their weblog become very introspective.

The primary purpose of Sigmund is to let the user discover the conceptualisations derived from the statistical data in a user-friendly way. The main functionality is:

1. Discover the conceptualisations in a single weblog. This is achieved by showing the persons mentioned, the (compound) terms used and the co-occurrence networks of persons and terms (see Figure 1 for an illustration).
2. Term-by-term comparisons. The co-occurrence method computes a value for each pair of terms and these values can be used to establish "agreement" between bloggers.
3. Compare multiple weblogs. Informally we say that two or more weblogs share a conceptualisation if the network of terms overlap.

Figure 1 shows a screenshot of Sigmund user interface. There are four browsers at the top. The left most browser shows the weblogs included in the analysis. When the user selects one, the other three browsers are filled with the names of persons mentioned in the selected blog and the terms extracted respectively.

Terms in the three browsers are sorted by descending frequency. The distinction between the browser labelled “absolute” and “relative” is that the former contains the absolute frequency of the terms and the latter the relative frequency taking into account whether terms are part of more compound terms. For example, the word sequence *knowledge management* can be seen as the word *knowledge* followed by the word *management*. The “absolute” browser counts 1 for *knowledge*, *management* and *knowledge management*. The “relative” browser counts only 1 for knowledge management. The user can select which of the two points of view is taken. Generally the “relative” measure produces more interesting results as it better reflects the way people use language to express concepts.

The user can explore the co-occurrence between an entry in one of the browsers with the other terms found in the selected weblog. This results in a network like the one shown in the lower part of the Fig. 1. Networks are generated as follows:

1. First, all terms that co-occur with the selected term above the threshold, or co-occurrence degree, are displayed.
2. Next, all displayed terms are compared with each other and if the co-occurrence degree exceeds the threshold a connection is drawn.
3. Note that the selected term that triggered the network to be generated is not visibly connected to the other terms for reasons of readability.

In Sigmund the co-occurrence degree (Section 3.3) is scaled to be between 0 and 100 for practical reasons, see for example Fig. 5.

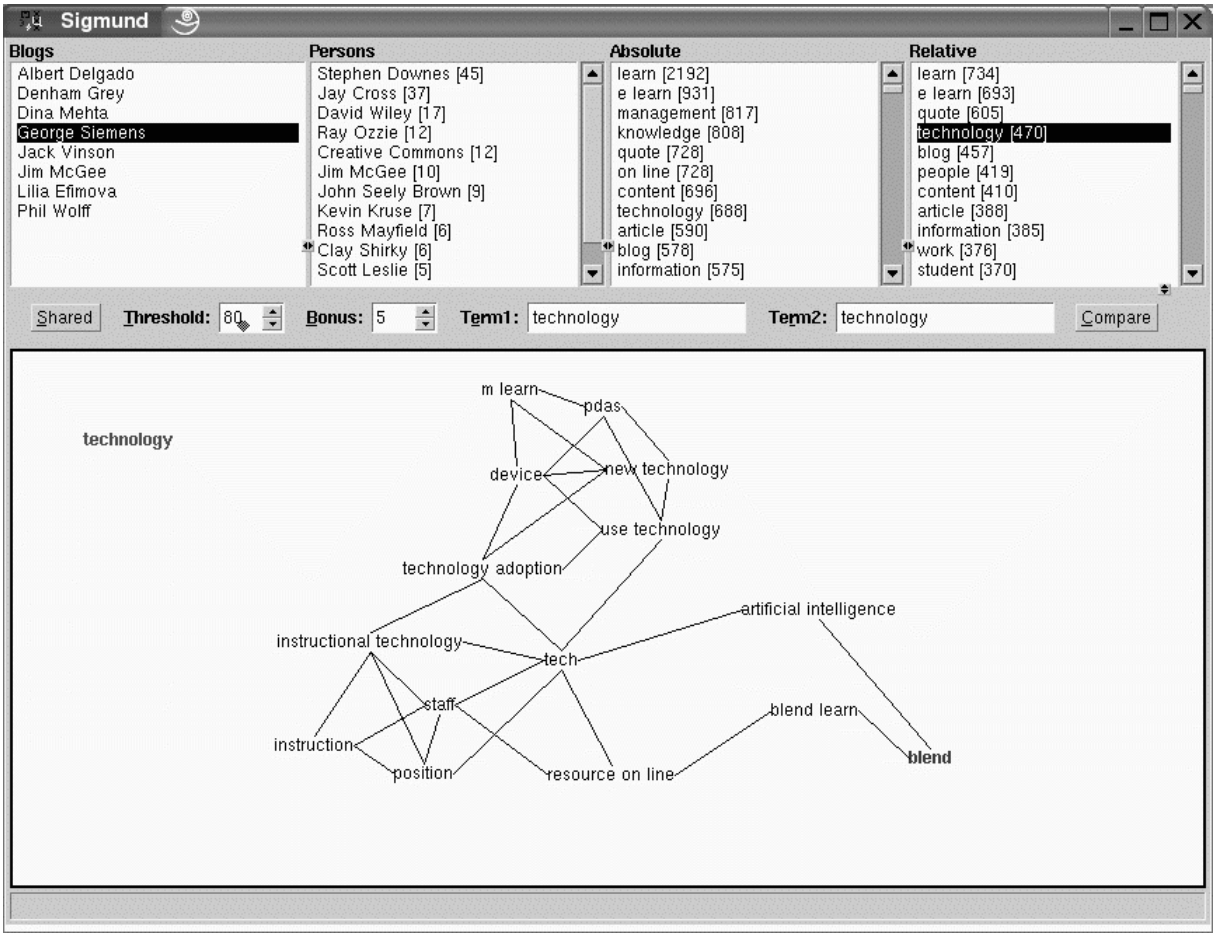

Figure 1. Sigmund user interface

5. Experience

In this section we illustrate Sigmund using examples from analysing weblogs in the knowledge management domain. Eight weblogs were included in the analysis. This list is limited by our ability to locate KM weblogs, some limitations of our spidering technique, as well as our choice to only consider weblogs that are exclusively in English. The current selection should suffice by way of illustration.

5.1. Bird's eye view on a weblog

The first group of Sigmund functions provides a bird's eye view on what a weblogger is talking about: a list of names, and absolute and relative frequencies of the terms used in the weblog. Fig. 2 illustrates how the tool could be used to provide a fingerprint of what a person is talking about: other people and terms. It shows the results of an analysis of a weblog maintained by one of the authors of this paper, Lilia Efimova. As people and terms are sorted by descending frequency these lists give an insight of the social network and the domain of her weblog. For example, the list of terms indicates that she writes a lot about weblogs and blogging, learning, people, knowledge and knowledge management.

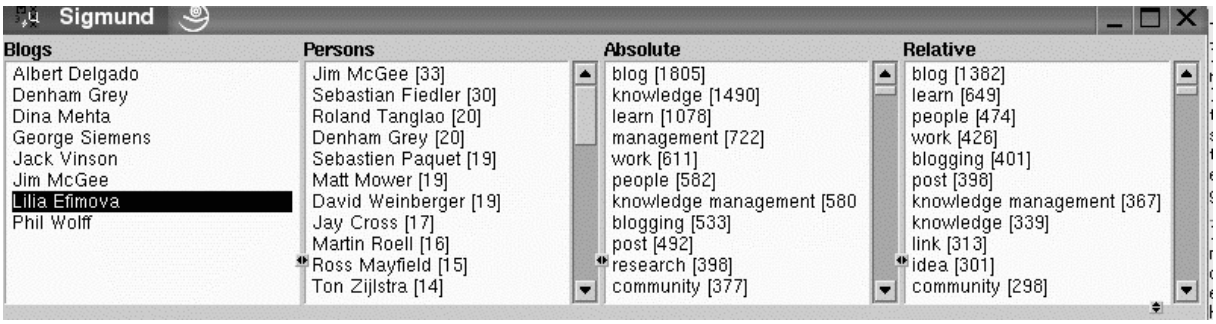

Figure 2. The most frequent terms in Lilia Efimova's weblog.

As these lists are automatically generated and provide an aggregated view of a weblog they could be contrasted with the opinion that bloggers, and possibly readers, have of their weblog. In the case of Lilia, the list of the names looked a bit surprising: her current collaborators and the people she considers influential were not listed as high as she expected.

There are at least two reasons for this. First, there is a limitation of the tool: at the moment it is capable of locating persons only if their full names are mentioned at least once and are unambiguous. Many bloggers have the practice of referring to bloggers they know well using only their first name (Nilsson, 2003). In this case names of people are usually accompanied with links to their weblogs, providing an unambiguous identifier for each person. Currently links are not included in our analysis.

Of course, this is not the only possible explanation of the differences between the generated lists and the expectations of the blogger. The social network of a blogger is not necessarily the same as the network of people that influence her thinking, nor is it necessarily the network of the most avid bloggers. People who influence writing do not always appear in the text of a weblog (they may be linked from a blogroll; see Marlow (2004) for differences between links in a text and blogrolls). Finally, someone's subjective view on what a weblog is about could be different from the more "objective" picture emerging as a result of analysing a weblog by tools like Sigmund.

Next to providing a high-level view of a weblog, the tool allows zooming in and analysing conceptualisations for a specific term or name (Fig. 3 and 4).

Fig. 3 provides an example of the co-occurrence network for a selected term (*knowledge management*) by another blogger, Jim McGee. It shows terms co-occurring with the term *knowledge management* and co-occurrences between these terms. It includes, for example, terms that could be classified as KM actors (*chief knowledge officer*, *individual knowledge worker*, *employee*, *partner*) and KM processes (*capture* and *share knowledge*).

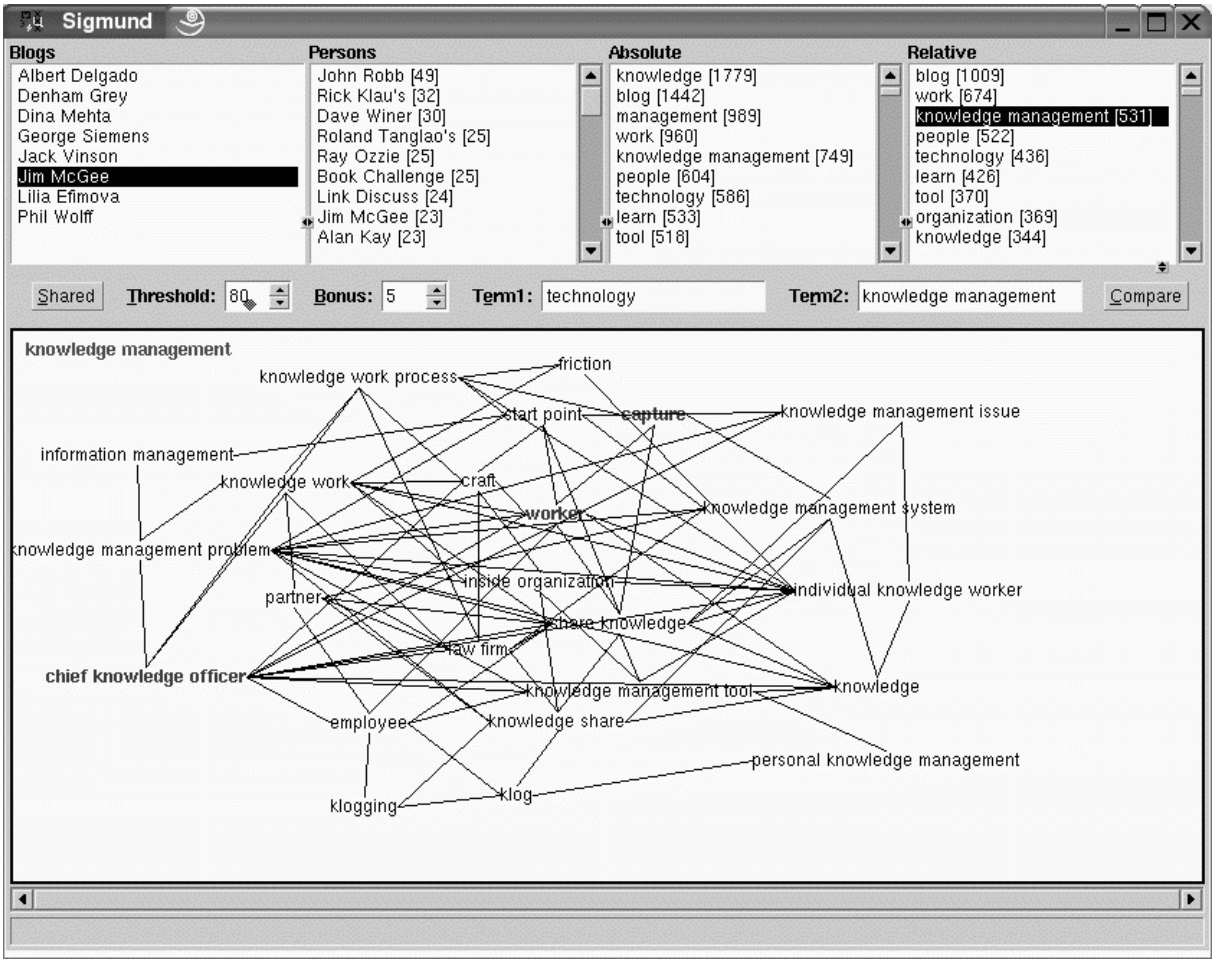

Figure 3. Co-occurrence network of *knowledge management* for Jim McGee's weblog

This figure reveals some interesting connections. For example, *knowledge work*, *knowledge work process*, *individual knowledge worker* and *chief knowledge officer* are co-occurring with each other, while *friction* is only connected to the first three. According to our subjective judgement these co-occurrences "correctly" identify a semantic relationship. Likewise *Sharing knowledge* is co-occurring with *knowledge management problem*, *knowledge management issue* and *starting point*, while *capturing* is not, which is in correspondence with Lilia's views on the matter. It is also apparent that some terms that may be considered as synonyms (e.g. *share knowledge* and *knowledge share*) are grouped together (e.g. *worker* and *employee*, *knowledge management issue* and *knowledge management problem*).

Fig. 4 shows another type of conceptualisation, a network of terms co-occurring with the name of a specific person. The figure shows that Lilia's posts that mention Jim McGee are mainly co-occurring with *knowledge work*, *knowledge workers* and *personal knowledge* (and it also indicates that they are likely to share preferences for *full text RSS*).

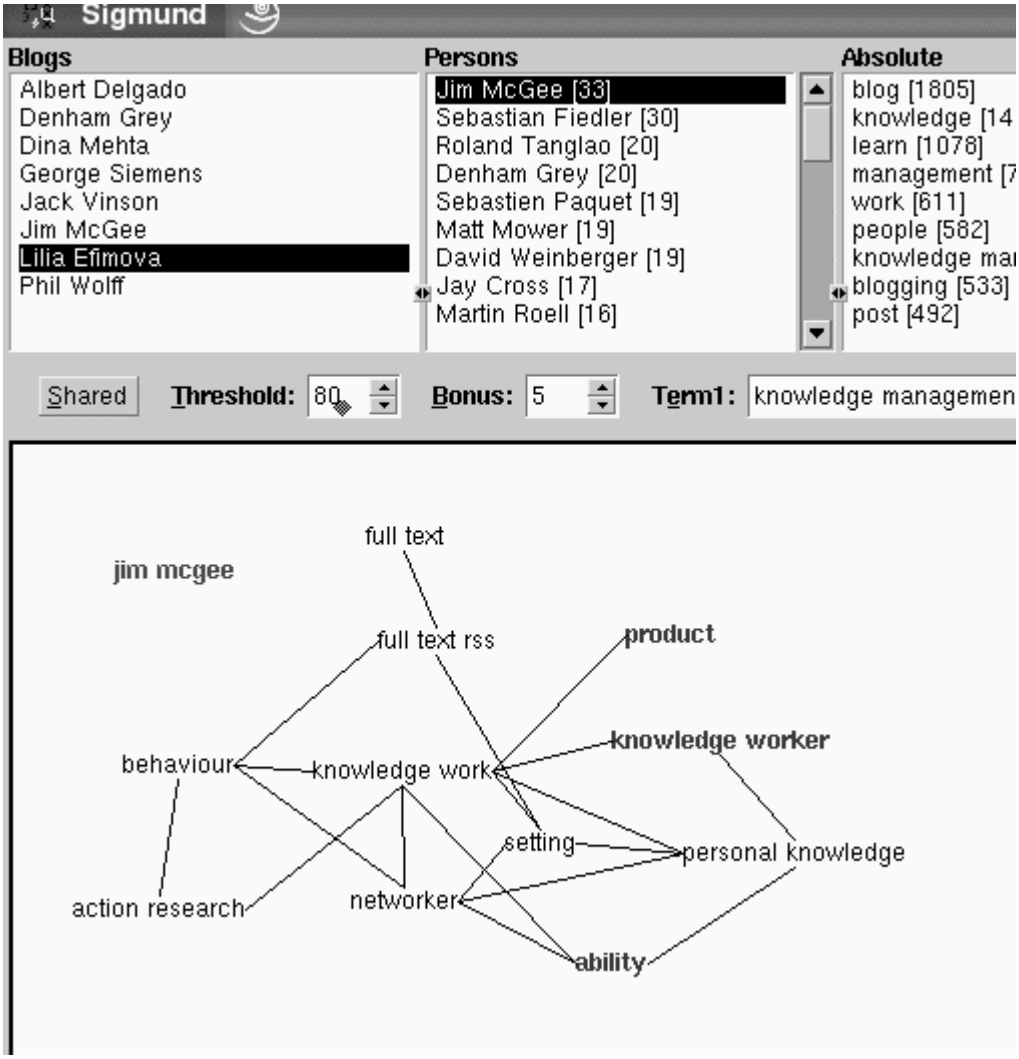

Figure 4. Co-occurrence network of *Jim McGee* for Lilia Efimova's weblog

Figures 2-4 provide examples of possible uses of Sigmund. Using it one can get an overview of a specific weblogs, and also explore the context of the usage of specific terms or names in it. Exploring the co-occurrence networks as clues for conceptualisations gives an insight in the mental models that guide someone's writing and can provide a better understanding of a text. Similarly co-occurrence networks with a given name provide context for the relation between bloggers and may point to domains of influence or conversations.

5.2. Comparing weblogs

The second group of functions allows to compare views on the co-occurrence between two terms for all weblogs, as well as finding how far co-occurrence networks of two or more bloggers overlap.

Fig. 5 shows co-occurrences between *knowledge management* and *software* in all weblogs loaded into Sigmund. For both terms there is an indication of their frequency in the text and a bar showing (with length and colour) how likely one term appears in a weblog post given the another term.

Blogger	Freq	knowledge management	software	Freq
Albert Delgado	47			63
Denham Grey	136			6
Dina Mehta	64			53
George Siemens	340			165
Jack Vinson	174			19
Jim McGee	531			151
Lilia Efimova	367			34
Phil Wolff	104			118

Figure 5. Normalised co-occurrence degrees for *knowledge management* and *software*

In the given set of weblogs, Denham Grey is most likely to write about *software* when he uses *knowledge management*, while Dina, George, Jim and Lilia do not mention *software* next to *knowledge management* often. When writing about *software*, Albert and Dina mention *knowledge management* often, but Lilia and George rarely.

As stated above (Section 3.3) the tool visualises co-occurrences of terms, so a strong connection means that a blogger uses terms together frequently, but we should carefully distinguish this from a specific relation between them. For example, Denham writes frequently that *knowledge management* is NOT about *software*, but about dialogue and social networking. Our current methodology does not make this visible.

The tool has a limited ability to recognise synonyms automatically. For example, Fig. 5 illustrates only co-occurrences between *knowledge management* and *software*, while some bloggers can use *tool*, *technology* or *system* when talking about software.

Fig. 6 (see next page) provides an example of comparing co-occurrence networks between two weblogs, those of Lilia Efimova and Jim McGee which gives some insight in their shared conceptualisations.

This figure represents 17 clusters of shared terms. Some of these indicate simple and obvious semantic relations. (e.g. *question – answer*, *movable type – software*, *knowledge work – knowledge worker – worker*). Others are not so obvious, but seem to be a point of agreement between Lilia and Jim (e.g. *content management system – marketing – corporate blogging* or *document – information overload – news aggregator*). Especially interesting is a large network of terms in the middle: it shows terms from several domains and co-occurrences between them. It includes a dense *education – learning* cluster, *community – social network – social software* cluster and *social capital – tacit knowledge* cluster.

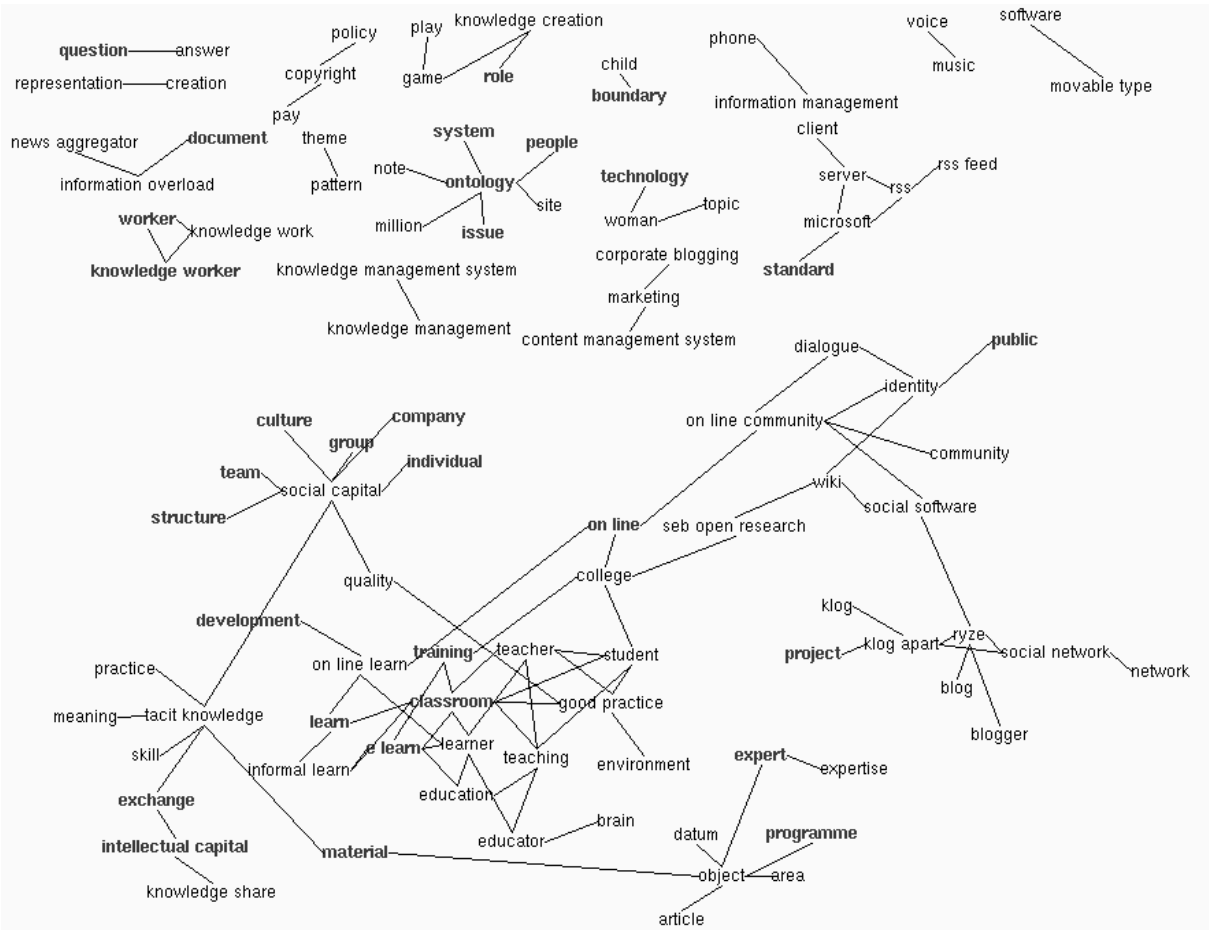

Figure 6. Overlap between co-occurrence networks between weblogs of Efimova and McGee

This visualisation indicates that while both authors write about knowledge management (both weblogs are recognised as KM weblogs) they seem to have more agreement on related areas (e.g. learning and education) or specific KM sub-domains (e.g. communities or social capital) rather than sharing conceptualisations about knowledge management itself. Only a few small clusters from the complex network of KM-related terms of Jim (Fig. 3) are left in this picture once we start looking for overlap with conceptualisations of Lilia.

6. Discussion

The ideas in this paper and their implementation in Sigmund was inspired by closely following the emerging field of weblog research and the call for proposals of BlogTalk 2.0. Some of the ideas were developed at the BlogWalk 1.0 workshop, the concluding dinner and, using transcendental methods, the social event.

For the implementation we have mainly relied on *tOKo* (Anjewierden et al., 2004) a toolkit for semi-automatically extracting ontologies from corpora of unstructured texts such as on-line forums and e-mail archives.

A fundamental motivation of our research is addressing the question of how the abundance of knowledge and insights available in digital form can be accessed. Traditionally, this problem has resided in the realm of Information Retrieval (IR) where the solution relies on posing the "right" query and human scanning of the documents returned. IR has been a spectacular success and is considered to be the most practical approach as evidenced by the popularity of search engines like Google. However, the use of IR has the severe drawback that people searching for information can only retrieve it in terms of the (almost) exact wording of the

original author rather than in more abstract terms. Thus IR is most useful if one has a clear idea of what to find. Moreover, even if a searcher knows what s/he is looking for, scanning documents to separate the wheat from the chaff is a non-trivial exercise. On the other end of the spectrum there are the efforts of researchers on the Semantic Web who seem to prefer to see the world structured *a priori* in terms of ontologies, to which documents can be linked. This approach has the major benefit that authors can be explicit about their intentions, an example is the use of RSS in weblogs. However, as may be obvious, linking millions of documents cannot practically be done manually and several papers and tools address the issue of automating this process (Pérez and Macho, 2004).

Our approach is an attempt to address the problem of operationalising the very idea of a higher semantic level. The most direct but crudest approach to finding the meaning of documents is indexing its terms. This is the realm of IR. Our basic assumption is that the conceptualisation of the meaning of documents is found as much in the way concepts are organised together as they are in the terms themselves, like the properties of molecules are determined as much by the arrangement of the constituent atoms as they are by the kind of atoms themselves. Finding co-occurrences seems to be the simplest approach for detecting (possible) relations between terms that can bootstrap *ab initio*, even without a sophisticated language model. Weblogs have proven a fruitful ground for this approach because they are both readily available and have the benefit of being naturally subdivided in posts with a single main focus. At a later stage we hope to make good use of the many links found in blogs.

An equally important motivation for the current research was to discover and operationalise and help knowledge flows. Knowledge flows and learning seem to be associated with one individual or group picking up the conceptualisations of the others. Thus, it is important to know what those conceptualisations actually are. We cannot warn enough that the current tool and methods only visualise co-occurrences of whatever happens to be written in the particular weblogs under analysis and that further interpretation for those suggestive networks is needed. However, with those warnings they do seem to give a glimpse at the conceptualisations of the authors and the shared understanding that is (or equally important) is not there.

7. Conclusions and Future Work

The current implementation of the tool is exploiting the graphical programming environment provided by SWI-Prolog (Wielemaker and Anjewierden, 2004) as well as the language technology part of tOKo (Anjewierden et al., 2004).

The largest practical problem we encountered is the availability of a weblog spider as outlined in Section 3.1. Our current version of such a spider is reasonably reliable, but by no means perfect. The only persistent solution appears to be that blogging software provides a public interface with access to all posts of a blog in a preferably established format such as RSS. Obviously, weblog research in general would greatly benefit from this.

We plan to make Sigmund and the weblog spider available publicly (see the weblog of the first author for details). This will, we hope, stir suggestions for improvements and additional functionality as well as trigger further research and applications.

Acknowledgements. This work was partly supported by the Metis project (http://metis.telin.nl). Metis is an initiative of the Telematica Instituut, Basell and Océ. Other participants are CeTIM, TNO, University of Delft, University of Amsterdam, University of Tilburg and University of Twente (all in The Netherlands).

Bibliography

Esma Aïmeur, Gilles Brassard and Sébastien Paquet (2003). *Using personal knowledge publishing to facilitate sharing across communities*. M.Gurstein (Ed.), Proceedings of the 3rd International Workshop on (Virtual) Community Informatics: Electronic Support for Communities - Local, Virtual and Communities of Practice. Available at http://is.njit.edu/vci-www2003/pkp-sharing-across-comms.pdf

Anjo Anjewierden, Bob J. Wielinga, and Robert de Hoog (2004). *Task and domain ontologies for knowledge mapping in operational processes*. Metis Deliverable 4.2/2003, University of Amsterdam.

R. H. Baayen, Richard Piepenbrock and Leon Gullikers (1995). *The CELEX lexical database (release 2)*. CD-ROM. Linguistic Data Corporation, University of Pennsylvania, Philadelphia.

Wayne Daniel (1995). *Biostatistics*. Wiley, 6th edition.

Asunción Gómez Pérez and David Manzano Macho (2003). *A survey of ontology learning methods and techniques*. OntoWeb Deliverable 1.5, Universidad Politécnica de Madrid.

Thomas R. Gruber (1993). *A translation approach to portable ontologies*. Knowledge Acquisition, 5(2):199-220.

Susan C. Herring, Lois Ann Scheidt, Sabrina Bonus, Elijah Wright (2004). *Bridging the gap: A genre analysis of weblogs*. Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04).

Cameron Marlow (2004). *Audience, structure and authority in the weblog community*. Presented at the International Communication Association Conference, New Orleans, LA. Available at http://overstated.net/media/ICA2004.pdf

Bonnie A. Nardi, Diane J. Schiano and Michelle Gumbrecht (2004). *Blogging as social activity, or, would you let 900 million people read your diary?* Submitted to CSCW 2004. Available at http://home.comcast.net/~diane.schiano/CSCW04.Blog.pdf

Stephanie Nilsson (2003). *The function of language to facilitate and maintain social networks in research weblogs*. Umea Universitet, Engelska lingvistik.

David W. Sabo (2003). *Relative risk and the odds ratio*. Section 8 in *Probability and statistics for the biological sciences*, British Columbia Institute of Technology. Available at http://www.math.bcit.ca/faculty/david_sabo/apples/math2441/toc2003.htm

Ariel S. Schwartz and Marti A. Hearst (2003). *A simple algorithm for identifying abbreviation definitions in biomedical text*. In: Proceedings Pacific Symp. Biocomputing, Kauai, Hawaii.

Stefan Staab and Rudi Studer eds. (2004). *Handbook on ontologies*. Springer-Verlag.

Jan Wielemaker and Anjo Anjewierden (2004). *Programming in XPCE and SWI-Prolog*. University of Amsterdam. http://www.swi-prolog.org