

# Detecting Knowledge Flows in Weblogs

Anjo Anjewierden<sup>1</sup>, Robert de Hoog<sup>2</sup>, Rogier Brussee<sup>3</sup>, and Lilia Efimova<sup>3</sup>

<sup>1</sup> Human Computer Studies Laboratory, University of Amsterdam, Kruislaan 419, 1098 VA Amsterdam, The Netherlands, [anjo@science.uva.nl](mailto:anjo@science.uva.nl)

<sup>2</sup> Faculty of Behavioural Science, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands, [R.deHoog@edte.utwente.nl](mailto:R.deHoog@edte.utwente.nl)

<sup>3</sup> Telematica Instituut, PO Box 589, 7500 AN Enschede, The Netherlands, [{Rogier.Brussee,Lilia.Efimova}@telin.nl](mailto:{Rogier.Brussee,Lilia.Efimova}@telin.nl)

**(Submitted, please contact first author for citation information)**

**Abstract.** Scanning the internal and external environment of an organisation to trace current and emerging topics of discourse, is an important task for knowledge management. Results of this scanning process can be used to anticipate on new developments and for detecting possible knowledge bottlenecks. This paper addresses the unobtrusive identification of conceptual structures exchanged and possibly shared in weblogs as a case to explore the opportunities for automated support for this scanning process. The research is also motivated by the observation that many domains defy formal conceptualisation in the sense that professionals, in particular knowledge workers, differ on the meaning of the concepts and terminology and the relations between them. In such domains, we use knowledge management itself as an example, detailed formalisation is pointless and the best we can aim for is finding out whether professionals agree and share, or disagree.

Informally, we define a *knowledge flow* to be the communication of some knowledge (represented as a natural language post on a weblog in our case) to a receiver (a reader of the post in question) when the receiver acknowledges the flow by referring back to the sender in a separate post. The research challenges we pursue are to determine what knowledge is being exchanged and to what degree the participants in the communication share conceptualisations.

We have implemented the identification of knowledge flows in a tool called BlogTrace. BlogTrace uses semantic web technology to represent weblogs and linkage structures, term extraction to determine concepts in natural language posts, and a statistically founded algorithm which computes a relevance metric for the concepts extracted and their relation to other concepts for any particular weblog author.

Based on several assumptions, software was written that had to answer knowledge management relevant questions concerning the frequency, intensity and topics of knowledge flows in a community of bloggers, which act like members of an actual organisation. The software was tested on a set of questions and the results show that this software provides to a large extent the answers.

## 1 Introduction

Professional knowledge workers, for example in the areas of law, journalism, computing and knowledge management, are increasingly using weblogs (blogs for short) to exchange and share information in their area of expertise. Blogs provide a low cost, easy to use, and more or less informal webpublishing platform. Although blogs are often viewed as “personal journals”, and have been compared to diaries, knowledge workers succeed in using them as an effective tool to create *spontaneous* virtual communities in which observations are passed on, questions are answered and issues discussed.

A weblog can be viewed as a collection of posts. Each post has a precise hyperlink on the web called the permalink, a title, and a date published. The permalink, which is a

simple URL, makes it possible for other bloggers to link to a post and thereby comment or reflect. Posts are generally self contained and permalinks, technically and socially, are essential to keep the blogosphere stable.

One of our research interests is in conceptualisations and more specifically: can we derive conceptualisations from (large) bodies of natural language text? Weblogs provide an ideal data source for doing this kind of research. Because of the nature of blogs we know they are written by a single person, which provides an opportunity to derive personal conceptualisations. By comparing conceptualisations of different bloggers we can obtain a metric of overlap and, implicitly, difference. From a broader, application oriented, perspective the ability to derive and compare conceptualisations from (semi)public sources in an unobtrusive way, can be very useful for organisations that have to scan the internal and external environment for tracing and detecting current and emerging topics of discourse. In an actual organisation the involvement of individuals in this discourse can be of help for knowledge management to stay informed about the waxing and waning of topics of interest that arouse the attention of its knowledge workers.

In this paper, we propose the idea of a *knowledge flow*. In the context of weblogs a knowledge flow could take place when a post of one blogger links back to a post of another blogger. Such a link could, of course, be accidental or trivial (“Have a look at ...”). It could also be the instance of a genuine exchange of knowledge between the bloggers involved. Our hypothesis is that the likelihood of the latter is larger when the overlap of conceptualisations between the bloggers is present in the terms used in the linked post. This hypothesis is an intuitive operationalisation of the assumption that using terms frequently and knowing what they mean are closely related.

We have implemented the idea of detecting knowledge flows in a tool called Blog-Trace. With the implementation of the analysis of knowledge flows it becomes possible to retrieve pairs of linked documents that are about a similar topic.

The paper is organised as follows. Sect. 2 provides a theoretical background for the idea of knowledge flows. Sect. 3 describes our approach to automatically extract conceptualisations from text documents and relates it to formal concept analysis. Sect. 4 presents the RDF/OWL based architecture to represent weblog data and, finally, Sect. 5 discusses the results of applying the knowledge flow idea based on the tools and methods described in the previous sections.

## 2 Communicating Knowledge, Action and Context

Basically, a knowledge flow is a specific type of information flow. From communication science [9], [5] we can borrow the sender-message-receiver model to describe these flows. The sender and receiver are points in a communication network, which exchange messages. Who the sender and who receiver is, depends on who takes the initiative. The sender-message-receiver model can be rephrased by the following three questions:

- Who says
- What
- To whom

In its most simple form detecting and analysing knowledge flows entails identifying the “Who” and “Whom” and the nature of the “What”.

For knowledge flows in weblogs, the easy part is the “who” part. By definition it is the person who is the author of the weblog.

We can define our way out of the “to whom part”. Many people can read a blog and be influenced by it without ever leaving a trace. However, as we will define a knowledge flow by visible actions in the form of links, quotes and comments only, we can (cheaply) define the reader as the person who writes these.

The tricky part is the “What”. If we strictly adhere to Shannon’s original sender-channel-receiver model that originated in designing error correcting communication channels there is no real problem. Unfortunately this model assumes that there is very little context that the sender and receiver have to share and agree upon. In the original model it is restricted to some alphabet and an encryption method. However, for human communication many “messages” can be understood only if the implicit or explicit context is taken into account. This context can be the author’s physical, cultural or economical environment, but more mundanely previous anecdotes, people the author is in contact with or the web pages read. Thus determining which context the message is addressing is an important part of determining what the content of the message is. Fortunately it turns out that the public blog medium forces people to be more explicit about their context and the technical infrastructure makes this comparatively easy by the use of links and copy paste.

However, this still does not answer the question how we can distinguish between a “mere” information flow and a knowledge flow. Obviously this depends on what we mean with “knowledge” and how we can distinguish it from “information”, which could throw us back to epistemological questions about the definition of knowledge. To short-circuit this discussion, we will simply adopt the following definition, which seems to be supported by the majority of the literature:

*Knowledge is a capacity to act* ([7], p. 37)

This definition implies that the content of the “What” will determine whether it is a knowledge flow or not, but also what (if anything) “Whom” is doing with the message. The one visible action that we can determine is what a reader puts in his or her blog. Thus we have defined a knowledge flow in terms of messages that refer to each other in terms of links or quotations and comments. At a closer inspection this only pragmatically solves the definition problem. The difference between information and knowledge still lingers.

Analysing knowledge flows in weblogs should serve some purpose. The information extracted from the data should be of help for answering questions. An area that could benefit from analysing knowledge flows in weblogs is knowledge management. Knowledge management is concerned with the knowledge household of a company, information must be about the knowledge household: its state, its development, its effectiveness and so forth, but also the external environment of an organisation. As this list can be extended, we have to choose a limited number of relevant areas in which knowledge flow information can be of value:

- Monitoring the *frequency* and *intensity* of crucial flows between people in and outside the organisation (the “Who” and “Whom”).

- Evaluating the content of the crucial flows between organisational entities to detect *bottlenecks* and emerging *problems/topics* (the “What”).
- Monitoring the development of *flows over time* to keep track of developments in the knowledge household.

These general problems can be made more specific as queries to be directed at a set of weblogs. The analysis software should be tested against these queries. This will be taken up in Sect. 5.

### 3 Extracting Conceptualisations from Text

In order to determine what the content of the information or knowledge that flows between the sender and receiver of the messages is we need to understand how the symbolic information in messages relates to the real world, to which we have limited access. Thus we think of the content of the messages as a representation of the real, or at least message external, world, *a conceptualisation*. The flow of knowledge is then the way the conceptualisation of the message sender is serialised in the conceptualisation in the message and the way the conceptualisation of the receiver is build up from the conceptualisation in the message.

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. Under the flag of ontologies ([6]) and the semantic web, it has captured a wider audience and led to knowledge representation languages such as RDFS and OWL. The approach aims to formalise structures that are believed to exist in some abstract sense in the real world, thereby making classification and inferencing possible.

In view of the above our aim is not to discover or formalise relations between concepts but to discover how bloggers (or other writers) describe “the world” without any *a priori* claim on the validity of those descriptions. The approach settles on finding patterns in the traces bloggers leave through their use of terms and linkage, and doing relatively crude statistical analysis of correlations. We assume that those correlations are a result of bloggers using an underlying conceptualisation. We *do not* assume that this conceptualisation is a faithful representation of the world, that it is shared between bloggers, or that the correlations are the underlying conceptualisation themselves.

If we accept that web presence is part of the “real world” we have at our disposal links and the use of orthographical equivalences of terms. In summary, the approach is as follows:

1. Identify terms and links 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 Sect. 3.1.
2. Once the concepts have been identified, we need to establish whether they are semantically related, at least according to a 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” (co-occurrence) of using one term given that blogger is using another term in the same post. See Sect. 3.2.
3. The output of the above is a two dimensional table of co-occurrences for using terms in combination. A typical weblog contains thousands of terms, so the table

contains millions of entries. We then select the high co-occurrences 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 Sect. 3.3 for the methodology and Sect. 5 for examples.

### 3.1 Term and Name Extraction

This section briefly describes how we extract terms and names from weblog posts. A more detailed account can be found in [1]. The main challenges with respect to language technology are the identification of meaningful terms, the extraction of the names of people, and expanding abbreviations to their long form.

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 sequence of nouns. 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 [2], 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 form. Because of the noisy nature of weblogs traditional abbreviation finding algorithms 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 frequent. A stop list of very common abbreviations (e.g. PC, CD, OS, etc.) is used to prevent accidental expansions.

3. 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.

4. Names of people are recognised using a gazetteer of first names based on the one in GATE<sup>4</sup> to which we added all names of players from the list published by international

---

<sup>4</sup> <http://gate.ac.uk>

chess federation to get a better world-wide coverage of names and a gazetteer with person name prepositions such as *de* (Robert de Hoog) and *bin* (Osama bin Laden). The algorithm also disambiguates names if only the first name is used, a frequent practice in weblogs. This is partially based on names occurring in the anchor text of a link.

### 3.2 Co-occurrence Analysis and Concept Structures

Intuitively a term or link  $B$  co-occurs with a term or link  $A$  if the frequency of  $A$  in posts containing  $B$  is much higher than the frequency of term  $B$  in posts not containing term  $A$ . This is not symmetric in  $A$  and  $B$ : for example in a document on management, the term *knowledge* could co-occur with *management* if the phrase knowledge management would be an important source of occurrences of the word knowledge, but if the document contains many uses of the term management alone or in phrases like customer relation management, management would not co-occur with knowledge. We quantify co-occurrence using the following statistics.

**Definition:** Let  $n(B | A)$  (respectively  $n(B | \neg A)$ ) 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  $n(* | A)$  (respectively  $n(* | \neg A)$ ) 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**  $c(B | A)$  is defined as

$$c(B | A) = \frac{n(B | A)/n(* | A)}{n(B | \neg A)/n(* | \neg A)}, \quad 0 \leq c(B | A) \leq \infty$$

We say that  $B$  co-occurs with  $A$  to at least degree  $k$  if  $c(B | A) \geq k$ . Note that  $c(B | A) = 1$  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.

The co-occurrence matrices can be thought of as a statistical version of the formal concept lattice [4] defined by the “occurs in” relation between terms as intent and posts as extend. Given a stream of random posts  $\{P_1, P_2, \dots\} = \{P_i\}_{i \in I}$ , consider the substreams of posts (or paragraphs)  $\{P_a, P_b, P_c, \dots\} = \{P_i\}_{i \in I_A}$  containing  $A$  respectively  $\{P_x, P_y, P_z, \dots\} = \{P_i\}_{i \in I_B}$  containing  $B$ . If we now consider the formal concepts  $C_A = \langle A | P_{ii \in I_A} \rangle$  and  $C_B = \langle B | P_{ii \in I_B} \rangle$ , the co-occurrence  $c(B | A)$  is the compare factor of the rate in which  $B$  occurs in the subconcept  $C_{AB} = \langle AB | P_{ii \in I_A \cap I_B} \rangle$  versus the rate in which  $B$  does not occur in the subconcept  $C_{AB}$ . In the limiting case  $c(B|A) = \infty$ ,  $C_B$  does statistically not exist as a separate formal concept: only as the subconcept  $C_{AB}$  does. In the limiting case  $c(B | A) = 0$  there is no subconcept  $C_{AB}$  but only the bottom subconcept  $\langle AB \dots | \emptyset \rangle$ .

The previous discussion addresses only the top layer

$$\begin{array}{ccc} & \langle \emptyset | \{P_i\}_{i \in I} \rangle & \\ & / \quad \backslash & \\ \langle A | \{P_i\}_{i \in I_A} \rangle & & \langle B | \{P_i\}_{i \in I_B} \rangle \\ & \backslash \quad / & \\ & \langle AB | \{P_i\}_{i \in I_A \cap I_B} \rangle & \end{array}$$

of the formal concept lattice. Although we have not pursued this, we can strengthen the interpretation of the co-occurrence matrix as a statistical version of a concept lattice

as follows. Define  $n(B_1B_2\dots | A_1A_2\dots)$  as the number of occurrences of the *collection* of terms  $B_1, B_2, \dots$  in posts containing the collection of terms  $A_1, A_2, \dots$  and define the co-occurrence degrees  $c(B_1B_2\dots | A_1A_2\dots)$  similarly as above. Again  $c(B_1B_2\dots | A_1A_2\dots) = \infty$  means that there is no separate concept  $\langle B_1B_2\dots | \dots \rangle$  but only a formal subconcept  $\langle B_1B_2\dots A_1A_2\dots | \dots \rangle$  of the concept  $\langle A_1A_2\dots | \dots \rangle$ .

### 3.3 Detecting Knowledge Flows

A potential knowledge flow occurs when, in a weblog post, one blogger includes a hyperlink to a weblog post of another blogger. We hypothesize that the intensity of a knowledge flow between two linked posts depends on the terms used in these posts and the conceptualisations of both bloggers as determined by the co-occurrence algorithm described in the previous section.

The operationalisation for the knowledge flow intensity is based on the idea of comparing the co-occurrence degree  $c(B | A)$  of the terms appearing in the linked posts. To bootstrap the process we assume that the knowledge flow search process is initiated with a cue term  $A$ .

**Cue term.** The cue term  $A$  must appear in both posts.

**Shared terms.** We define a shared term to be a term  $B$  that occurs in both posts and for which both bloggers have a co-occurrence degree  $k$  such that  $c(B | A) \geq k$ . The informal interpretation of a shared term is that both bloggers associate it with the cue term on the basis of the entire blog.

**Agreed terms.** A term  $B$  that occurs in both posts and for which one blogger has a co-occurrence degree  $k$  such that  $c(B | A) \geq k$ .

**Private terms.** A term  $B$  that occurs in precisely one of the posts and for which  $c(B | A) \geq k$  holds for the blogger of the post.

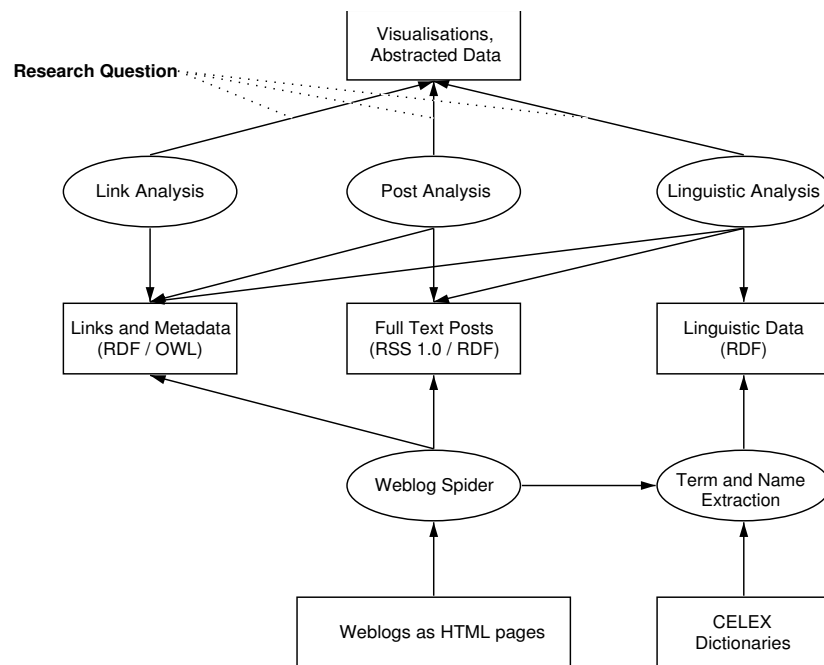
Experimental results are given in Sect. 5.

## 4 Data Acquisition and Representation

Although data on which algorithms described in the previous section could be applied is widely available. Most data is not in a format suitable for processing. This section describes the main functions of BlogTrace. BlogTrace is a software environment that defines ontologies to represent links between documents, documents themselves and linguistic information that can be used by the co-occurrence and knowledge flow algorithms.

A lot of understanding of weblogs comes from bloggers themselves often with generalisations based on personal experiences, observed blogging practices or an analysis of unrepresentative samples. However, for a researcher looking at the blogging phenomenon it is especially important to understand if it is representative for weblogs as a whole, for a specific blogging subculture or just for an individual blogger. This points to a need for a software environment that accomodates blog research. BlogTrace is such an environment.

We will briefly describe the architecture of BlogTrace as far as it is relevant for this paper (see Fig. 1). The most critical aspect is turning the HTML pages that represent a weblog into a more suitable representation through spidering and mining. This process



**Fig. 1.** Overall architecture of BlogTrace

consists of extracting metadata (name of the blogger, title of the blog), identifying the parts of the HTML that are posts and turning the post attributes (title, permalink URL, date and body) into metadata. Although weblogs are generated by software, there is a high variety in the layouts used. Our approach to the spidering and mining problem is to run induction algorithms on the HTML pages. These algorithms generate patterns (for example based on attributes in HTML elements and/or the position in the DOM-tree). A calibration algorithm then tries to find a stable set of patterns that fit the natural order of post attributes.

The output of the blog spider is represented as RDF using the semantic web library described in [8]. One file contains all metadata about the blog itself, all permalinks and all hyperlinks inside the posts. This data can be used to perform linkage and community type research at the blog level (e.g. [3]). The second file contains the full texts of the posts in RSS 1.0. This data can be used for analysis at the post level. The third file contains linguistic data abstracted at both the blog and post levels (terms, term frequency and in which posts these terms are used), see Sect. 3.1. The linguistic data supports the identification of conceptual structures, both for individual bloggers and for a community of bloggers and is fundamental for the knowledge flow analysis in this paper.

Although BlogTrace currently only works for weblogs, the spider looks for features only blogs have, there are no restrictions to apply it to other types of documents that have the same structure: text interspersed with links.

The metadata about a weblog is represented using class `docs:Weblog`, most properties refer to the Dublin Core<sup>5</sup> and FOAF<sup>6</sup> ontologies:

<sup>5</sup> <http://purl.org/dc/elements/1.1/>

<sup>6</sup> <http://xmlns.com/foaf/0.1/>

```

<docs:Weblog rdf:about="http://anjo.blogs.com/metis/"
  dc:creator="Anjo Anjewierden"
  dc:description="The source code is the ultimate documentation"
  dc:title="Anjo Anjewierden"
  rdfs:label="http://anjo.blogs.com/metis/"
  foaf:nick="Anjo Anjewierden">
  <docs:hasRssFeed rdf:resource="http://anjo.blogs.com/metis/rss.xml"/>
</docs:Weblog>

```

A hyperlink (HTML: `<a href="...">`) is represented using class `link:SimpleLink` from a link ontology. The `link:sourceDocument` is the document in which the hyperlink was found and `link:targetDocument` is the document to which the hyperlink points.

```

<link:SimpleLink
  link:anchorText="city metaphor to explain blogging">
  <link:sourceDocument
    rdf:resource="http://anjo.blogs.com/metis/2004/06/city_metaphor_f.html"/>
  <link:targetDocument
    rdf:resource="http://blog.mathemagenic.com/2004/06/07.html"/>
</link:SimpleLink>

```

Posts are represented using `rss:item` from the RSS 1.0<sup>7</sup> ontology. The `link:SimpleLink` above was extracted from the post represented as RSS below.

```

<rss:item rdf:about="http://anjo.blogs.com/metis/2004/06/city_metaphor_f.html">
  <rss:title>City metaphor for blogging</rss:title>
  <rss:link>http://anjo.blogs.com/metis/2004/06/city_metaphor_f.html</rss:link>
  <dc:date>2004-06-10</dc:date>
  <rss:description>
  <p>Another train discussion we had was Lilia's
  <a href="http://blog.mathemagenic.com/2004/06/07.html">city metaphor
  to explain blogging</a> post. [... ]
  </rss:description>
</rss:item>

```

Using the above data generated by the spider we can then exploit OWL to define higher-level notions. For example, in order to identify a knowledge flow we require links from weblog posts to other posts which can be defined in OWL as follows (in N3 notation):

```

link:WeblogPostLink rdfs:subClassOf link:SimpleLink;
  rdfs:comment "A WeblogPostLink is a SimpleLink if and only if
    both the source and the target documents are
    weblog posts (RSS items)";
  owl:intersectionOf (link:SimpleLink
    [ a owl:Restriction;
      owl:onProperty link:sourceDocument;
      owl:someValuesFrom rss:item
    ]
    [ a owl:Restriction;
      owl:onProperty link:targetDocument;
      owl:someValuesFrom rss:item
    ]
  ).

```

<sup>7</sup> <http://purl.org/rss/1.0/>



blogs, weblog posts inside a selected blog and links in a selected blog or post. The large area at the bottom is used to display results from end-user queries.

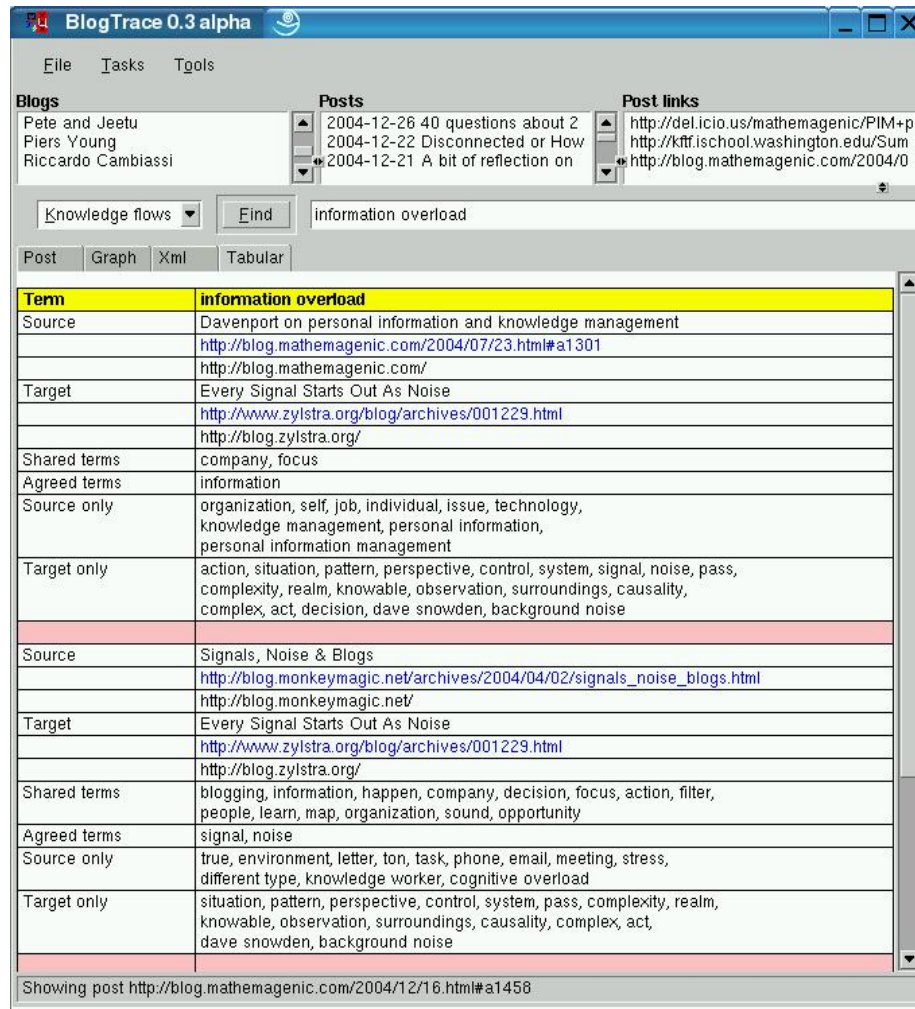
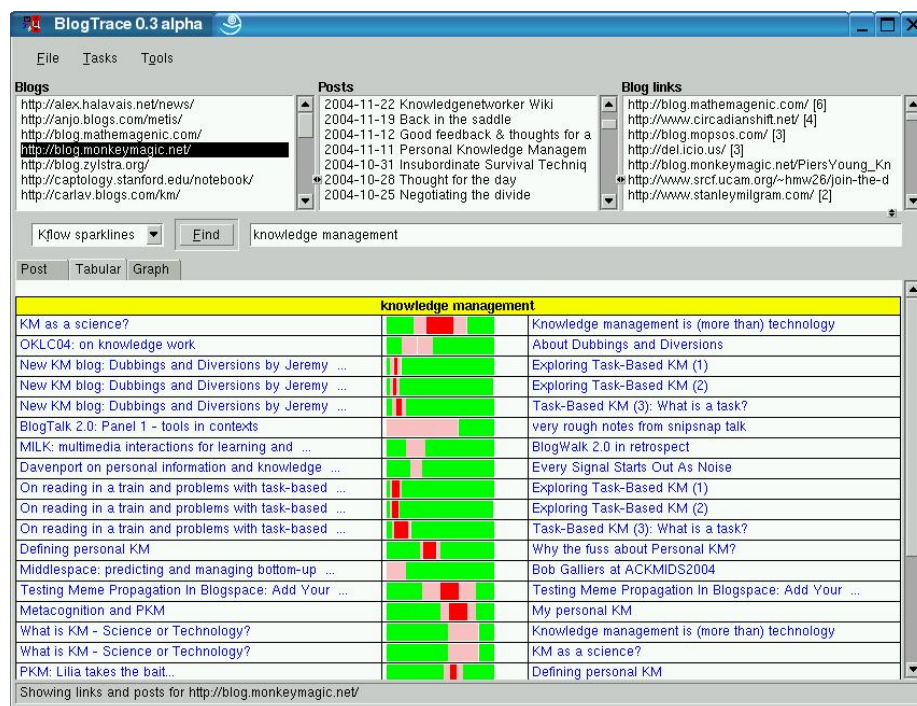


Fig. 3. Knowledge flows as terms

The concept  $C^*$  in this case is the term *information overload* entered in the search line in the middle part of the screen. This search term can be either free keywords or selected terms from a predefined vocabulary. The selected function is searching for *Knowledge flows*. The lower part of the window shows in a tabular form all the pairs of posts that are linked through *information overload*. The first part of the table shows the source and the target post and the weblogs to which they belong. The **Shared terms** (see Sect. 3.3 for the definition) are the terms both bloggers associate with *information overload*. The **Agreed terms** are terms the linked bloggers in these posts associate with *information overload* but do not belong to the shared terms. The **Source only** category contains terms only the source blogger associates with *information overload*. The same holds for the **Target only** category.

*Question 3:* Given an answer on Question 2, what is the amount of agreement or disagreement between bloggers about a concept? As was said already, initially we take the amount of term overlap as an indicator of the agreement and disagreement in a knowledge flow. Later we will try to analyse the content of the posts on specific terms that can indicate agreement or disagreement. The answer to this question is implicit in Fig. 3, but BlogTrace contains a convenient facility to make this more easy to see. By selecting the *Kflow sparklines* function after entering a search term, the tool presents the results using the idea of a *sparkline*<sup>8</sup>.



**Fig. 4.** Knowledge flows displayed as sparklines

In Fig. 4 the search term is *knowledge management*, to show that the software works for different concepts. Pairs of posts are depicted and the middle column gives a sparkline indicating the amount of overlap by using the following colour codes:

- **Shared terms** Shown in dark grey.
- **Agreed terms** Shown in light grey to the left and right of the shared terms.
- **Source/Target only** To either side of the shared/agreed terms in medium grey.

The smaller the dark and/or light grey bar the less overlap. The placement of the medium grey area gives an indication of the terms which are used only by one of the bloggers in their posts. By clicking on a post title both the linked posts are displayed in the lower window, enabling a closer inspection of their content.

<sup>8</sup> [http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg\\_id=0001OR&topic\\_id=1&topic=](http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR&topic_id=1&topic=)

The results described in this section show that BlogTrace can be used to analyse weblogs to detect knowledge flows in general, but also concerning specific concepts, between bloggers through cross-referencing in their posts. From a knowledge management perspective these results can be used to find answers for the kind of areas of interest mentioned in Sect. 2: monitoring the frequency and intensity of flows between people in- and outside an organisation, evaluating the content of these flows to detect bottlenecks (who agrees/disagrees with whom concerning what) and emerging problems and topics (agreed and/or target/source only terms).

## 6 Conclusions

In this paper we have investigated whether shared conceptualisations can be extracted from weblogs. Weblogs can be seen as an instance of a wider class of publicly accessible information sources which are bound to a specific individual. The notion of a knowledge flow was introduced to identify mutual concept dependent linkages between individual bloggers. Based on several assumptions, statistical methods and software was developed that had to answer knowledge management relevant questions concerning the frequency, intensity and topics of knowledge flows in a community of bloggers, which act like members of an actual organisation. The software was tested on a set of questions and the results show that this software provides to a large extent the answers.

During this process at least two important problems were detected at the technical level that need a solution in the future. **Quoting.** Quotes are used frequently in weblog posts. At the moment these quotes become part of the vocabulary of the quoter and thus incorrectly influence the conceptualisations derived. **Abbreviations.** The current method of processing weblogs cannot detect that in two *different* blogs shared abbreviations are used. As several concepts are often expressed as an abbreviation (for example PKM for “personal knowledge management”), important concepts can be missed.

Of course, the work reported here is only the start of a process that can lead to software that has a richer potential for analysing knowledge flows between weblogs and other types of documents. The following issues will be taken up:

1. The software does not make extensive use of natural language analysis. This limits the power of the analysis. On the other hand, full blown natural language analysis of weblogs is not the overall goal of the research. There should be some middle ground in which certain aspects of natural language processing can be used to improve the concept detection process and also the ability to identify agreement and disagreement at a semantically richer level than term overlap alone.

2. There are obviously other ways for individual weblogs to share conceptualisations than through explicit linking. One such way is a kind of indirect sharing when two bloggers don't link to each other but both refer to the same external link. An exhaustive list of direct and indirect referencing tactics used by bloggers shows that these can enhance the ability of the software to detect and analyse knowledge flows. A related issue is that of *weblog conversations* in which multiple bloggers link over time.

As a caveat one should keep in mind that the current analysis process is based on several assumptions that can be invalidated in the future.

**Acknowledgements.** This work was partly supported by the Metis project<sup>9</sup>. Metis is an initiative of the Telematica Instituut, Basell and Océ. Other partners are CeTIM, Technical University Delft, University of Amsterdam, University of Tilburg and University of Twente (all in The Netherlands).

## References

1. Anjo Anjewierden, Rogier Brussee, and Lilia Efimova. Shared conceptualizations in weblogs. In Thomas N. Burg, editor, *BlogTalk 2.0*, Vienna, July 2004.
2. R.H. Baayen, Richard Piepenbrock, and Léon Gulikers. The CELEX lexical database (release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 1995.
3. Lilia Efimova and Stephanie Hendrick. In search for a virtual settlement: An exploration of weblog community boundaries. Submitted, 2005.
4. Bernhard Ganter and Rudolf Wille. *Formal concept analysis. Mathematical foundations*. Springer-Verlag, 1999.
5. Claude A. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
6. Stefan Staab and Rudi Studer, editors. *Handbook on ontologies*. Springer-Verlag, 2004.
7. Karl-Erik Sveiby. *The new organizational wealth: Managing and Measuring Knowledge-Based Assets*. Berrett Koehler, 1997.
8. Jan Wielemaker, Guus Schreiber, and Bob Wielinga. Prolog-based infrastructure for RDF: Scalability and performance. In Dieter Fensel, Katia Sycara, and John Mylopoulos, editors, *2nd International Semantic Web Conference (ISWC)*, Sanibel Island, FL, USA, October 2003.
9. Sven Windahl, Benno Signitzer, and Jean T. Olson. *Using communication theory: An introduction to planned communication*. Sage Publications, 1992.

---

<sup>9</sup> <http://metis.telin.nl>