A. M. Heinecke, H. Paul (Hrsg.): Mensch & Computer 2006: Mensch und Computer im Struktur*Wandel*.
München, Oldenbourg Verlag, 2006, S. 313-320

Enhancing the meta structure of weblogs

Jochen Reich, Karlheinz Toni, Dr. Georg Groh

TU München, Lehrstuhl für angewandte Informatik

Abstract

Weblogging as a concept for publishing personal lifestyle as well as knowledge driven contents has gained popularity in the last years. In this paper we outline the advantages of weblogging and tool driven approaches to information gathering. We will introduce a solution to combine the advantages of both approaches by assigning a meta structure on weblogs.

1 Introduction

Many possibilities to satisfy information needs and support the knowledge worker have been introduced. Scientific work was strongly focused on organizational software systems [1]. In recent years the concept of weblogging has gained increased popularity. They appear not only as personal dairies but also as knowledge driven information spaces. They contribute to the knowledge management of profit and non-profit organizations. We will first compare structured information gathering as it is accomplished in the tool- based approach with unstructured information gathering as it is done in weblogs. The result is that weblogs have the potential to present an easy to use, dynamic and up to date knowledge base. What lacks is a meta structure, which would allow to perform extended information retrieval on a semantical basis. In the following chapters we will present how a meta structure can be assigned to weblogs.

2 Shortintroduction to Weblogs/Blogs

Weblogging is one of the latest concepts for publishing all kinds of recent information. [1] Although there is no general definition of the term weblog [3] some features are characteris-

[1] A general definition of the term information does not yet exist [2]. We define information as externalized knowledge. Knowledge derives from information which is interpreted by a human being.

tic and commonly accepted. A weblog is a website that contains information of users commenting about subjects of their fields of interest, e.g. other websites or happenings in their daily lives. As bloggers can publish contents immediately weblogs have the potential to contain up to date information on recent events. Weblogs are typically represented to the user as an ordered list in which the latest entries appear on top. Blogger can post links and cite or reply to entries [4]. We distinguish two kinds of weblogs; personal weblogs, containing exclusively private information and knowledge driven weblogs which aim at sharing explicit knowledge in profit or non-profit organizations. In this paper we will concentrate on knowledge driven weblogs as they deliver information assets in the term of acquisition of knowledge according to Probst [5].

3 Underlying scenario

In order to pursue the idea of knowledge management several approaches and theories have been introduced. A huge variety of knowledge management tools has been implemented [6][7], providing the knowledge worker with functionality to gain,structure, store, retrieve, share and provide different views of information. In this tool based approach the information acquisition is performed via predefined concepts, e.g. an entry mask. In this way the information and metadata can be stored withoutpostprocessing in a predefined format according to logical, semantical and syntactical rules. It can thus be easily retrieved by the tool. The knowledge management tool can to some extend verify and reason the information being provided by the user.

Weblogs on the other hand present an unstructured knowledge base. Metadata can be derived from the assumption about structural elements, e.g. the name of the header, and the linkage provides information about topically related entries. The main information is written in unstructured, natural language, in the body of the entry. The weblog can be seen as a semantic net in which the entries of the weblogs can be put into relation to each other. The relations themselves are specified by the textual description of the links provided in the entry. A meta structure is provided by the content itself, namely by the links. This meta structure is created by the user on subjective decisions and his own preferences. Even the topical relatedness of two weblogs connected by trackback or pingback links may be doubted. The reader may have missunderstood the weblog he refers to, oreven change the topic at will. The concept of weblogging lacks a meta structure which is complete, objective and helpful in performing information retrieval. The advantages of this approach is that because of the informal character and the provided simple user interface weblogging is easy to do even for users with average skills concerning the use of computers and no skills concerning web publishing. The openess of weblogging sites and the ease of use contribute to a dynamic, fast growing and up-to-date information base. The realization of a weblog is cheap, as only webspace and the weblog page itself with scripts enabling the blogger to publish his entries have to be provided. The weblogs can be globally accessed via a web browser.

The conclusion is that both approaches, structured and informal information gathering, have their advantages. The main advantage of the tool based approach is the predefined metastructure of the information base. Creating a comparable metastructure on weblogs would make

information retrieval easier and more sufficient on them. It will be possible to deliver a structured view of the information available according to the needs of the knowledge worker. We will introduce how such a metastructure can be builtin the next chapter.

The concepts introduced will also be applicable on Wikis as they share most of the characteristics of weblogs. Both weblogs and wikis can be interpreted as knowledge driven semantic nets whose contents can be collaboratively edited and enhanced.

4 Proposed solution

In this chapter we will introduce steps towards a structured weblog. The first step is to improve the linkage between the entries. To do so we have to extract the entries of the weblog. For the further treatment stop word removal and stemming are then accomplished. To establish associations between the entries their semantical distance is then computed. This distance is computed according to the vector space model under the consideration of the entries' structural elements. Determining the complete link structure is an information asset itself and represents the basis for the further ranked clustering of the entries. We will first subsume entries to clusters and compute their relevance within the clusters. The clustering is accomplished under the use of word vectors. The clusters represent only a unit of topical related entries but do not state the entries' relevance within the clusters. Thus our second step is to compute a entries relevance by applying the PageRank algorithm after having completed the linkage between the entries.

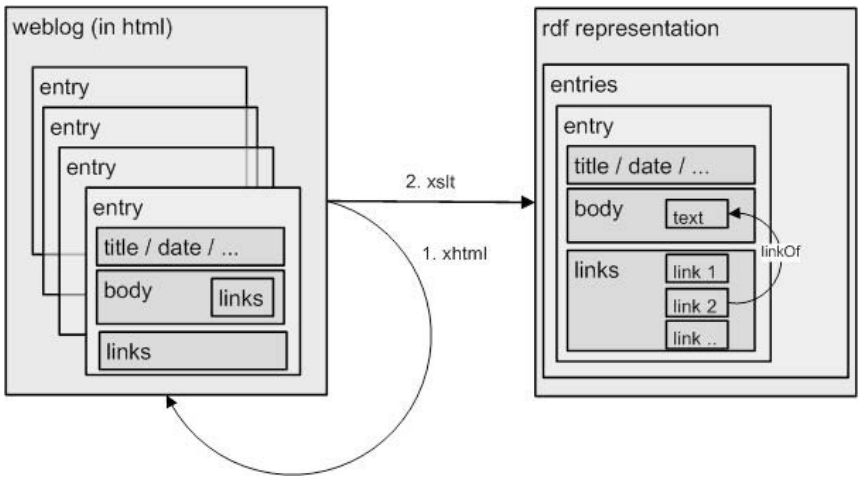

Figure 1: Extraction of a structural weblog elements to RDF

4.1 Extracting entries from webblogs

As we work on single entries of weblogs we have to extract these entries first. The weblog has to be parsed and recurring structural elements have to be examined as they potentially represent entries. A weblog from a technical view could for example consist of a table containing rows which represent the entries.

Once these recurring structures have been identified their content needs to be compared with characteristics of an entry of a weblog. Examinations of the number of words and links in one entry have been introduced in [8]. The authors estimate the typical length of a weblog as somewhere between 80 and 494 (mean = 209,3). They empirical found that a typical number of links is between 0 and 6 for internal links and 0 to 5,25 for external links[2]. An entry of a weblog typically does not contain images or other tables. The heading is at the beginning of the entry and is mostly highlighted by a bigger font size or color than the entry´s text. Oncerecurring structures are found that conform with the stated characteristics they can be considered as entries.

We implemented an easy to adapt "*SemBlog agent*" transforming the weblog, which is typically available as html, to xhtml. Our agent instantiates our OWL declaration by transforming the xhtml document via (automatically derived)[3] xslt to RDF. We do so in order to get a standardized, formal, easily accessible representation of the structural elements of the weblog (cf. figure 1).

4.2 Stop word removal

As we create a metastructure on weblogs in dependence of their semantics we only handle words that carry semantic information. As so called stop words appear typically in large numbers they are the dominating words in texts although they do not contribute to their content. Therefor frequently appearing conjunctions, punctuation marks and expletives will be removed. To do so a catalogue of stop words has to be available.

4.3 Stemming

In order to compute word vectors for clustering and applying the vector space model we need to overcome the problem of different variations of words. Although the words stopped and stopping are nearly semantically equal they differ in their orthography and may thus not be identified as the same word by text comparison algorithms. What is needed is a method to identify the root of a word. This procedure is called stemming. It is a highly language dependent problem.

Porter introduced an often cited approach which supplies transformation rules for stemming words in english language. The appraoch by Porter transforms a word step by step by deleting its prefixes and suffixes and afterwards reconstructing the end of the word. The deletion

[2] internal links point to resources on the same web domain, external links to external web domains

[3] Although the xslt is automatically derived from the xhtml document via the analysis of recurring structures, it still has to be revised manually

is done until a minimum number of syllables is reached. The resulting roots are not necessarily linguistical correct. But as the aim of stemming in the context of information retrieval is to deliver a logical view of words the results are sufficient. Implementations of this algorithm can be found easily in the web. It can easily be adapted to other languages by changing the transformation steps.

We have followed the suggestions in the evaluation of different kinds of stemming algorithms in [11]. Our SemBlog agent performs this step after stop word removal and structural normalization of the webblog entries (cf. section 1).

4.4 Computed and weighted interrelations

The entries of weblogs are connected by links the user sets. These links are equally from a technical view but carry different semantics. A blogger can set a link to an entry he replies to, cites, adds information to, or only points to. He can also link to resources outside the blog, somewhere in the web, which is called permalink. Other bloggers again can refer to his entry and create a so called trackback link. The linkage of weblogs is created by the bloggers on their subjective decisions, their interests and preferences. It has not the claim of being sound or complete. What is needed is a linkage structure that is complete and thus connecting all topical related entries. With this structure the user could be provided with a list of topically related entries and navigate through more entries than originally referenced by the one he was viewing. As most of the entries are related to each other even though to a small percentage a threshold has to be found above which they are significantly related. The user should be offered to specify this threshold due to his information needs. Also for the Pagerank algorithm which we will apply in chapter 2 the underlying link structure is of fundamental importance.

4.5 Consideration of structure

The only usable structural information a weblog provides is the determination between heading and the body of the entry. Headings contain only few words which typically outline the content of the following entry. Therefor their weight in comparing the topical relatedness of entries of weblogs is higher than the weight of the words appearing in the body of the entry. To meet this demand we multiply the number of words in the heading with a weight, automatically determined by our SemBlog agent. The weight depends on the mean semantic simularity of the single title words and the keywords in the body of the weblog entry as well as the mean semantic diffusion, i.e. whether the topical coherence is high or low[4].

4.6 Determining the semantical distance of entries

In the context of information retrieval in weblogs the vector space model poses an appropriate means for the computation of the semantical distance of entries. The linkage can then be accomplished according to the distance. The basic idea of this approach is to express the

[4]A paper about semantic diffusion and topical coherence, going into algorithmic detail is in preparation

content of texts as vectors and to evaluate their distance. For each entry e_k and the query q_l a vector is computed which will be compared. The dimension of the vector space is determined by the number of words of the underlying vocabulary V. It consists of positive, real valued weights. R^V mit $V = V(\{dk\})$. The weights can be determined according to the relative number of occurrences of a feature in the single entry or in the set of available entries. The feature is a term or part of a term (n-gram) [9] to which a dimension is assigned to. The mapping of the features of a given text to a vector is called *indexing*. The semantic of each entry and the querry is represented by the associated vector. The closer their distance the better the entry suits the querry. The distance is computed according to the following formula [10]:

$$R(d_k, q_l) = \mathrm{sim}(d_k, q_l) = \cos(\sphericalangle(d_k, q_l)) = \frac{q_l \bullet d_k}{\|q_l\| \cdot \|d_k\|}$$

$$= \frac{\sum_{t=1}^{V} w_{tl} \cdot w_{tk}}{\left(\sum_{t=1}^{V} (w_{tl})^2\right)^{\frac{1}{2}} \cdot \left(\sum_{t=1}^{V} (w_{tk})^2\right)^{\frac{1}{2}}}$$

In the context of weblogs the vector space model can be used for the computation of the association between topically related entries. Every entry is characterised by its vector[5].The vectors will be compared with each other by assigning the role of the querry vector to every vector one after another and applying the space vector model. The user could be presented a topic map showing to which percentage entries are related. Another view of the computed result could be a ranked list of the related entries.

4.7 Overlapping Clustering

To find groups of topical related entries we build overlapping clusters. As entries of weblogs strongly differ in the number of words a metric has to be found that takes this assumption into account. We suggest that if a word represents more than a user defined percentage of the entry´s text it can represent the name of one of the clusters. This is accomplished for every entry. Each cluster has a number of entries assigned to it. As an entry can contain several words which represent more than a given percentage of the text it can be assigned to several clusters. The clusters can thus be overlapping. The user could be provided with an index of available collections of entries.

4.8 Relevance of entries

What we have achieved in the previous chapters was to find sets of entries with the same topic which are all connected by a diverse number of links. When the user is presented the entries of a cluster a mean to indicate the relevance of each entry within the cluster is helpful. Consider a set of entries connected through links. The more entries refer to an entry the more

[5]note that the number of words appearing in the heading is weighted

relevance it has within this topic. Entries which are often linked to will be thus ranked higher than others.

4.8.1 Requirements for the evaluation algorithm

As weblogs are mostly unsupervised information spaces the algorithm for computing the relevance of entries would have to immediately or after a short deceleration evaluate their content. The algorithm will be applied to large amounts of entries. Duo to performance reasons it has to converge after a well defined and reasonable number of iterations. The number of entries will encrease and thus the algortihm has to be highly scalable.

4.8.2 PageRank

We suggest the PageRank algorithm as a possibility to state the relevance of entries of weblogs. New to our approach is that the algorithm will be applied to several parts of a web page. We will shortly introduce the concept of PageRanking. The main idea is that if an entity is referenced by many others its content is of high quality and thus deserves a high rank. Also the rank of the entities referencing this entity is of importance. Thehigher their rank the higher the rank of the referenced entity will be. Before applying the algorithm some entities have to be ranked manually. The less references an entity contains the higher is their value. As the PageRank consists of only a few calculation steps it is performant enough even if applied to large amounts of entities. In this context the entities are the entries of weblogs. The PageRank algorithm is as provided in the following listing.

$$R_i = (1 - d) + d(Pr(R_j/C_j)) \text{ mit } j = 1....n$$

R_i stands for a floating point number which represents the value on which the entries are ranked. The higher the number R_i the higher is the relevance of the entry *i*. R_j represents the trustworthiness of page *j*, which in this equation is devided by the number of links on this page *C*. This quotient is computed for every entry linking to entry *i* and the resulting values are added up. This sum is multiplied by an damping factor *d* with $0 <= d <= 1$ and *(1-d)* is added. The multiplication with the damping factor and the addition of *(1-d)* is needed to lead the equation to convergence. Convergence is reached after a certain number of iterations depending on the distance between the value of the initial relevance and its final value according to PageRank.

5 Application scenario

The initial point was that we were only able to perform full-text search on weblogs. The assets we have gained through our approaches will be pointed out in this chapter. Once the links to the topical related entries are computed we can provide the reader with different views of the semantical neighborhood of the entry he is currently reading. A map of topical related entries would show him every for his information needs relevant entry at a glance. The degree of relatedness and trustworthiness could be indicated by graphical elements and colours. To supplement this map and a full-text search a filter can be implemented that only returns entries above a certain degree of relevance specified by the user. In the map only

related entries above a threshold again specified by the user would be shown. If the reader requires an overview of the topics available in a weblog an index can be created. He could be provided a list of available topics. Chosing one he will be presented a complete list of the entries related to the topic sorted by their relevance.

References

[1] Lehel, V.; Matthes, F.; Steinfatt, K. (2003): Weblogs als ein innovatives Instrument des betrieblichen Wissensmanagements, Mensch und Computer 2003: Interaktion in Bewegung.Stuttgart: B. G. Teubner.

[2] Bick, M. (2004): Knowledge Management Support System – Nachhaltige Einführung organisationsspezifischen Wissensmanagements. Universität Duisburg-Essen.

[3] Westner, M. K. (2004): Weblog service providing. UNITEC Institute of Technology.

[4] Burg, T. N. (2003): Zum neuartigen von Weblogs, MONSTER MEDIA.

[5] Probst, G. J. B.; Gibbert, M. (2005): Strategic management in the knowledge economy. Publicis.

[6] Reininghaus, A.; Minrath, H.: Eureka: Wissensmanagement im technischen Kundendienst bei Xerox, Xerox.

[7] ARIS Process Platform, 2003, IDS Scheer AG.

[8] Nardi, B. A.; Schiano, D. J.; Gumbrecht, M.; Swartz, L. (2004): "I'm Blogging This" A Closer Look at Why People Blog. ACM Press.

[9] Wikipedia; 2006; http://de.wikipedia.org/wiki/NGram Analyse.

[10] Groh. G. (2001): Applying Text Classification Methods to the Mapping of simple Extensional Ontologies for Community Information Management, 2001, Universitaet Kaiserslautern.

[11] Hull, D.A. (1996): Stemming Algorithms: A Case Study for Detailed Evaluation Journal of the American Society of Information Science. 47, 70-84.

Kontaktinformationen

TU München
Lehrstuhl für angewandte Informatik
Jochen Reich

Boltzmannstr. 3
85748 Garching
JochenReich@in.tum.de