On the Conceptual Tag Refinement

Nenad Stojanovic
FZI at the University of Karlsruhe
Haid-und-Neu Strasse 10-14
76131 Karlsruhe
+ 49 721 9654 852
Nenad.Stojanovic@fzi.de

Ljiljana Stojanovic
FZI at the University of Karlsruhe
Haid-und-Neu Strasse 10-14
76131 Karlsruhe
+ 49 721 9654 804
Ljiljana.Stojanovic@fzi.de

Jun Ma
FZI at the University of Karlsruhe
Haid-und-Neu Strasse 10-14
76131 Karlsruhe
+ 49 721 9654 812
Jun.Ma@fzi.de

ABSTRACT
Social tagging is a well known approach of assigning keywords to (web) documents in order to share the personal meaning of users about the content. However, the ambiguity of the assigned keywords hinders the knowledge sharing process. In this paper we present an approach for supporting tagging process by interpreting the keywords (tags) using a conceptual Tag model, which leads to a semantic tagging process. Moreover, the approach introduces the tag refinement process that proposes extensions of given tags in order to help the user to better express his/her "tagging" need. We present several evaluation studies in order to demonstrate the efficiency of the approach.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

General Terms
Management, Performance

Keywords
Semantic Tagging, Conceptual Model, Tag Refinement, Ontology Pruning

1. INTRODUCTION
Social tagging has become a very useful method for gathering knowledge about a domain in a light-weighted manner using so called collaborative community portals (e.g. flickr or del.icio.us). However, due to a weak representation model (set of keywords), this collaborative knowledge engineering approach results in a weakly structured "tags cloud" that informally presents the current domain. The main problem is the ambiguity of used tags, so that they can be processed/combined only on the syntactical level. For example, the frequency of the appearance of a term in tags (in total) determines its importance, without considering the contexts in which the term appears in particular tags.

In this paper we present a novel method for enabling an efficient tagging process. The approach is based on the interpretation of the tags using a predefined conceptual model for tagging (so called Tag model), which leads to a semantic tagging process. Moreover, the approach supports users in the tagging process by providing them with the context of the tags, i.e. with the set of terms relevant for understanding what does a tag mean, including the lexical variants of these terms (e.g. synonyms). By considering that context, the user can easily extend/refine defined tags. The approach is based on the work in the query refinement [1]. Indeed, we treat this challenge for tagging using the experience from the information retrieval (search in particular): very short queries are ambiguous and an additional refinement step is required in order to clarify what a user meant with a query, i.e. what is his information need. In an analogous way we are talking about the annotation (tagging) need, i.e. what did the user consider as relevant in a document that he has tagged: if the tag consists of only 1-2 keywords, then it is very difficult to conclude properly what the user meant with that tag/s. Consequently, the usage value of given tags decreases. For example, due to this ambiguity in the meaning, tags are not very suitable for the development of the conceptual model of the target domain (but they could be used). Therefore, our approach introduces conceptual tag refinement as a method for the disambiguation of the meaning of tags given by a user, with the goal to support the user in the tags' extension (i.e. providing more metadata about a document).

We have developed an annotation framework that realizes this idea and we have used it for annotating web documents in three use-case studies from the eGovernment domain. The evaluation results are very promising and we present some results related to the quality of the tag refinement process.

The paper is organised as follows: In the second section we present our approach for the conceptual tag refinement. In the third section we give some implementation details and present evaluation results. In section four we present some related work and in section five we give some concluding remarks.

SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

2. CONCEPTUAL TAG REFINEMENT

2.1 The problem
The main problem in understanding tags associated to an information source is the ambiguity in the interpretation of the meaning of tags. We can consider here the analogy to the well-elaborated "problem" of short Boolean queries in the information retrieval: due to the ambiguity in the interpretation of the meaning of a short query, lots of irrelevant document are retrieved. Similarly, keyword-based tags do not define in a disambiguous way the intension of the user in the tagging process. For example, by tagging a page with terms "process" and "knowledge", a user has open a large space of possible interpretations what he meant: 1) "process knowledge" as a type of knowledge, 2) "knowledge

process" as a type of process, or 3) "process that is somehow, but not directly related to knowledge".

Therefore, the main problem for the effective usage of tags is to determine the context in which the user has used them in order to know what the user meant by using these tags. A solution is to transform a set of tags (several keywords), that a user has defined in the tagging process for a document into a more structured form, that will enable better understanding of the user's intention in the tagging process. Such a method will be elaborated in the rest of this section.

2.2 The Conceptual Tagging Model

In order to enable better interpretation of a tagging (i.e. tags, that a user has assigned to a document) we have defined the Conceptual Tagging Model, illustrated in Figure 1. This work is based on the work in the Conceptual query refinement [1].

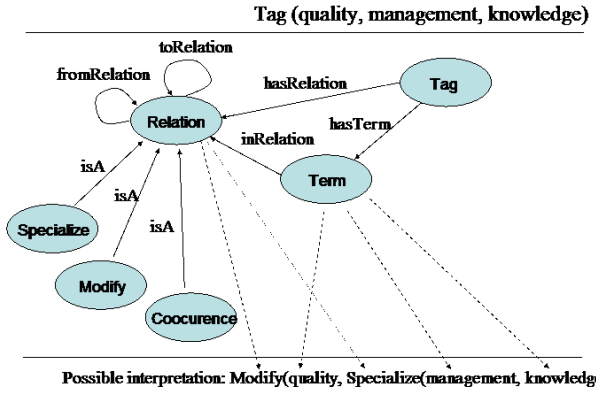

Figure 1. Conceptual model for interpreting the meaning of a tag is represented inside the central grey box. An example of interpreting the tag "quality, knowledge, management" is illustrated at the bottom

The interpretation of a tag depends on the interpretations of the individual tag terms (cf. Figure 1, concept Term). The interpretation of a tag term can be defined through its relations with other terms (concept Relation, cf. Figure 1), whereas a relationship can be established between concepts Relations as well. Therefore, a meaning of a tagging is represented as a context that encompasses nested relations between tag terms (i.e. relations between relations and terms are allowed too). A meaning of a tag represents so called tagging (annotation) need. For example, if a user has defined the tag = "knowledge, management, quality" for a document, then the meaning (i.e. tagging need) "quality regarding knowledge management" can be in the given model represented as a relation between the term "quality" and a relation between terms "knowledge" and "management", like rel2("quality", rel1("knowledge", "management")). This is called a Tag model. Note that a tag might have several meanings, i.e. it can be mapped into several Tag models.

We introduce several types of relation, which a term belongs to. Indeed, from the conceptual point of view, two terms are either in a specialization relation (cf. Figure 1, Specialize) (i.e. a term specializes the meaning of another term, like "process + workflow = process workflow"), in a modification relation (cf. Figure 1, Modify) (i.e. a term modifies the meaning of another term, like e.g. "process executed by a workflow") or in a co-occurrence relation (cf. Figure 1, Co-occurrence) (i.e. two terms just appear together in a context, without a direct influence on each other, like e.g. "… Process is described using a model, which can be found in the literature about workflows"). For the completeness of these relation set we rely on the previous work in the query conceptualisation [1].

2.3 Building Tag Models

As we already mentioned, a meaning of a tag (i.e. its interpretation) is defined through the relations that can be established between tag terms, i.e. for a tag (t_1, t_2, … t_n) a meaning is defined as $rel_{x1}(t_1, ... (rel_{xz}(t_{n-1}, t_n))\ldots)$, whereas $rel_{xi} \in \{$Specialize, Modify, Cooccurrence$\}$.

Since a user defines a tag regarding an information resource, an interpretation of the tag should emerge from that resource. On the other side, the main information that can be derived from an information resource is the linguistic one (shallow NLP): there is a verb, there is a noun, there is a noun phrase, etc. Therefore, in order to build the meaning of a tag, one has to process this information and derive conceptual relations between tag terms. Figure 2 sketches the process of building tag models.

buildQueryModels

input:

Tagging: t_1, t_2, ... t_n

Document *D*

output: ordered list of Tag models:$rel_1(t_1, rel_2(t_2, ...))$,

$rel_i \in \{$*Specialize*, *Modify*, *Cooccurrence*$\}$.

1. Find the set of sentences from the document D that contain one of query terms (*SetSen*)

2. Perform shallow NLP on the set *SetSen*

3. Find the set of noun phrases from *SetSen* that contain one of tagging terms (*SetNP*)

4. Determine the structure (relation) in which each tagging term appear in the context of another query words, regarding the set *SetNP*

5. Create Tag models

Figure 2. Procedural description of building query model. Some more explanations are given below in text

DT(JJ)*(NN)*, DT(JJ)*(NNS)*, DT(JJ)*(NP)*, DT(JJ)*(NPS)*, DT(NN|NNS)*, DT(NP|NPS)*, (JJ)*(NN|NNS)*, (JJ)*(NP|NPS)*, PP$ (JJ)*(NN|NNS)*

whereas:

DT	determiner, general (a, the, this, that)
JJ	adjective, general (e.g. near)
NN	noun, common singular (e.g. action)
NNS	noun, common plural (e.g. actions)
NP	noun, proper singular (e.g. Thailand, Thatcher)
NPS	noun, proper plural (e.g. Americas, Atwells)
PP$	pronoun, possessive (e.g. my, his)

Figure 3. Linguistic patterns for finding noun phrases (an excerpt)

In the nutshell of the approach is the consideration that the relations between the noun phrases determine the relations between tag terms contained in them. As we already mentioned, structural information are derived from the linguistic processing based on the shallow NPL performed in step 2 (cf. Figure 2). Figure 3 and 4 represents several examples of the linguistic patterns used in the steps 3 and 4, respectively. Complete description of linguistic patterns can be found in [1].

NPx: Noun Phrase
NP1 => Spezialize(a, b), whereas a ∈ NP1 and b ∈ NP1 and previous(a, b)
NP1 of|at|with|from|on|in NP2 => Spezialize(a, b), whereas a ∈ NP1 and b ∈ NP2
NP1 for NP2 => Modify(a, b), where a ∈ NP1 and b ∈ NP2
NP1 VP NP2 => Modify(a, b), where a ∈ NP1 and b ∈ NP2
(VB|VBD|VBN|VBG|VBZ)* => VP: Verb Phrase
VB verb, base (believe)
VBD verb, past tense (believed)
VBG verb, -ing (believing)
VBN verb, past participle (believed)
VBZ verb, -s (believes)
Cooccurency relation exists between two terms that co- occurre in the same sentence but which are not neither in a Specialize nor a Modify relation

Figure 4. Linguistic patterns for detecting relations (an excerpt)

Therefore, by mapping the conceptual space to the linguistic space we can build tag models.

2.4 Using background lexical knowledge

The conceptual tag refinement derives semantics from statistical and structural information, whereas the structure is determined by NLP (Natural Language Processing). However, due to the usage of various vocabularies, valuable statistical information is lost. Through aggregation of the semantically identical terms (synonyms) the plausibility of a refinement can be increased consequently. In this work we use WordNet as a valuable source for such relations.

Without going into details, we are mentioning here two main usages of the WordNet:

1) Synonyms from WordNet are used as an extension in searching for linguistics patterns. As an example we define such an extension for the first pattern from Figure 4:

NP1 => Spezialize(a, b), where (a ∈ NP1 or (ax=synonym(a) ∈ NP1)) and (b ∈ NP1 or (bx=synonym(b) ∈ NP1)) and before(a, b)

2) Hypernym relation can be used for a) defining Specialize relation and b) extending the linguistics pattern in the same way as defined for synonyms.

2.5 Tag refinement

By introducing the presented conceptual model of tagging, the tag refinement process can be seen as the refinement of a Tag model. Indeed, for a tag (q_1, q_2, … q_n), where q_i i=1, n are tag terms, several Tag models can be built and for each of them several refinements can be generated. For example: tag(t_1, t_2, … t_n) has several meanings in the form $rel_{x1}(t_1, ... (rel_{xz}(t_{n-1}, t_n))\ldots)$, that can be refined in several ways, e.g. $rel_{x1}(t_1, ..., rel_{xz-1}(t_{n-2}, rel_{xz}(t_{n-1}, t_n)))\ldots)$

whereas $rel_{xi} \in$ {Specialize, Modify, Cooccurrence}.

Therefore, by using a conceptual representation of a tag, a refinement is not represented just as a bug of words, but rather as a structure with an implicitly represented meaning. It will enable a user to better understand the refinements and to focus on semantic relevant ones.

Table 1: Linguistic patterns for finding tag models in the text. ***NP(C1)*** means Nounphrase containing C1. ***NP(C1C2)*** means Nounphrase containing C1 and C2, C1 must appear before C2.

Tag model	Possible Meaning		Sentences patterns	Result examples
Specialize(C1,C2)	**Single noun phrase**	NP(C1C2)	"(.)*" + ***NP(C1C2)*** + "(.)*"	*The future must be a new* ***computer world*** *with high speed.*
	Double noun phrase	NP(C2) IN NP(C1)	"(.)*" + ***NP(C2)*** + ***Inphrase*** + ***NP(C1)*** + "(.)*"	*It is not enough to change the real* ***world*** *with the* ***computer*** *science.*
Modify(C1,C2)	**Single noun phrase**	NP(C2C1)	"(.)*" + ***NP(C2C1)*** + "(.)*"	*We are producing the* ***world*** *famous* ***computer*** *.*
	Double noun phrase	NP(C1) VP|VIN NP(C2)	"(.)*" + ***NP(C1)*** + ***verb phrase pattern*** + "|" + ***Vinphrase pattern*** + ***NP(c2)*** + "(.)*"	***Computer*** *is being produced to the* ***world*** *market.*

3. IMPLEMENTATION & EVALUATION

The approach presented in the previous section has been implemented in an annotation tool, used for semantic tagging of three eGovernment organisations' web portals. This work has been performed in the scope of the EU IST project FIT, http://fit-project.org. The main role of the tagging process was the support of the ontology pruning process, as illustrated in Figure 5.

The first task in the approach is the ontology learning process, which an ontology is developed in (c.f. Figure 5).

In the second step the traditional tagging process is performed, i.e. a user should define several keywords that express his/her view on a document (or a part of the document).

In the third step, the contextualisation of the tags is performed: the tags defined by users are put in the context of 1) the text, they are summarizing and 2) already existing domain. The result is the Tag model that formalizes the meaning of the user's tags, by defining the relations between tags and their context.

In the last step the previously derived Tag Model is mapped onto existing ontology in order to define the most appropriate position (if any) for including/placing the user's tag in the ontology, which is out of the scope of this paper.

In order to demonstrate the usefulness of the proposed approach we have performed several evaluation studies. For example, the quality of the developed ontology was evaluated in [2].

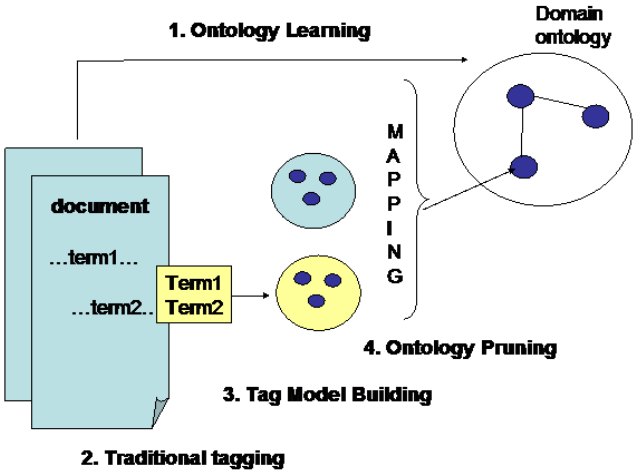

Figure 5. Main steps in the proposed approach for using conceptual tagging in pruning an ontology

In this paper we evaluate the quality of the tag refinement process. The approach is evaluated on the web documents taken from the three previously mentioned eGovernment organizations. The evaluation was user-driven. We selected ten users, who are familiar with the chosen domain (public administrators from the corresponding organisations). The task was to tag (at least) twenty pages by using the proposed approach (i.e. the annotation tool presented in section 3). We measured the quality of the suggested refinements for each particular tagging given by users (200 taggings in total) by calculating the relevance of refinements (Table 2) and the precision of the ranking of these refinements (Table 3).

Table 2: Calculated precision and recall for suggested refinements (defined separately for one-word and two-word tags and the suggested Specialize and Modify refinement)

Tagging contains:	Suggested Specialize relations		Suggested Modify relations	
	precision	recall	precision	recall
one word	95%	91%	93%	80%
two words	90%	86%	91%	82%

Table 3: Calculated precision@5 for selected refinement by a user, i.e. in how many cases the refinement taken by a user has been placed in the first 5 refinements suggested by the system

Tagging contains:	precison@5 for Specialize relations	precison@5 for Modify relations
one word	85%	80%
two words	75%	62%

Finally, Table 4 illustrates the efficiency of the approach by showing the structure of consuming time in the tag refinement process.

Table 4: Average time spend in two phases of the approach

Tagging contains:	Linguistic processing	Tag refinement
one word	250ms	450ms
two words	400ms	600ms

Discussion: Although middle sized, these evaluations have shown the reliability of the approach for practical applications. The man problem is that the some relevant refinements cannot be found (c.f. Table 2, columns "recall") – the main reason is the weakness of the NLP approach that has been used in this work (Text2Onto has been used [3]). The precision of the generated refinement is quite high, which means that the system produces useful refinements. The ranking approach (c.f. Table 3) can be improved, but it can be done only by using more contextual information about the current user. Finally, Table 4 shows that by improving the efficiency of the linguistic processing, the approach will increase the performance for about 30%. It can be done either by using another processor or by more efficient integration of this phase in the whole process.

4. RELATED WORK

This paper presents a novel approach for formalizing social tagging process. Though some blueprints do exist, to our knowledge there has been no prior explicit formulation of this approach, nor a concrete application, as presented in this paper. Nevertheless, a substantial amount of related work already exists concerning the general goal of creating annotations for Web pages, as well as keyword extraction and in the following we discuss some of the most important works in these research areas.

Text data mining is one of the main technologies for discovering new facts and trends about the currently existing large text collections [4]. There exist quite a diverse number of approaches for extracting keywords from textual documents. In this section we review some of those techniques originating from the SemanticWeb, Information Retrieval and Natural Language Processing environments, as they are closest to the algorithms described in this paper.

In Information Retrieval, most of these techniques were used for Relevance Feedback [5], a process in which the user query submitted to a search engine is expanded with additional keywords extracted from a set of relevant documents [6]. Some comprehensive comparisons and literature reviews of this area can be found in [7].

Keyword association is useful for enriching already discovered annotations, for example with additional terms that describe them in more detail. Two generic techniques have been found useful for

this purpose. First, such terms could be identified utilizing co-occurrence statistics over the entire document collection to annotate [8]. In fact, as this approach has been shown to yield good results, many subsequent metrics have been developed to best assess "term relationship" levels, either by narrowing the analysis for only short windows of text [9], or broadening it towards topical clusters [10], etc.

Our approach differentiates from all of related work by the introduction of a conceptual model for the underlying extraction/tagging process. By introducing a kind of "conceptual distance", it enables more formal analysis of the extracted terms / proposed annotations, which leads to the possibility to define the context in which they can be added in the ontology. Other methods are based on simple/syntactical co-occurrence and cannot be successfully used in the ontology development process.

Finally, provision of semantically rich, ontology-based metadata is one of the major challenges in developing the Semantic Web. In recent years, various annotation systems have been developed to face this challenge. There is, however, a lack of systems that: 1) can be easily used by annotators unfamiliar with ontologies, 2) are able to annotate a part of a page; 3) gives advices for improvement. In this paper we presented our annotation tool aiming to satisfy these needs. It is browser-based in order to support wide and distributed usage. It has simple user interface that hides complexity of ontologies from the annotator. Finally, it alleviates the ontology development process.

5. CONCLUSION

In this paper we presented an approach for formalizing tagging process by interpreting the keywords (tags) using a predefined conceptual model (ontology), that leads to a kind of semantic tagging. The approach enables the disambiguation of the meaning of tags given by a user and therefore supports better knowledge sharing. Moreover, based on this Tag model, we introduced the tag refinement process, that recommends extensions of given tags in order to help the user to express his/her "tagging" need. In that way we stimulate the generation of the metadata related to an information resource. By knowing that this metadata is formally represented (machine understandable), it is clear that the importance of such an approach is crucial for the Semantic web development.

We have developed a tool for supporting the whole approach. The tool is realised as an annotation editor, which allows domain experts to tag interesting Web documents with their own personal concepts (folksonomies). By designing this tool we did not try only to exploit the advantages of these processes (easy to use vs. formal semantics). More important we tried to mix them in order to resolve their drawbacks. As already known, tagging systems suffer from having too little formal structure and easily result in "metadata soup". On the other side, annotation user interfaces cannot be based upon a closed, hierarchical vocabulary (i.e. ontology), since they are awkward and inflexible.

The tool has been applied in three uses cases in the eGovernment domain. The first evaluation results are very promising: the approach can support an efficient annotation process. However, the approach depends on the performance of the used linguistic processor and the efficient definition of the refinement patterns is an ongoing work.

Acknowledgement

The research presented in this paper was partially funded by the EC in the project "IST PROJECT 27090 - FIT".

6. REFERENCES

[1] N. Stojanovic, "Method and Tools for Ontology-based Cooperative Query Answering", PhD thesis, University of Karlsruhe, Germany, 2005

[2] L. Stojanovic, N. Stojanovic, J. Ma, An approach for combining ontology learning and semantic tagging in the ontology development process: eGovernment use case, WISE 2007 - the 8th International Conference on Web Information Systems Engineering, in press

[3] P. Cimiano, J. Völker, Text2Onto -A Framework for Ontology Learning and Data-driven Change Discovery, Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, pp. 227-238. Springer, Alicante, Spain, June 2005.

[4] M. A. Hearst. Untangling text data mining. In Proc. of the 37th Meeting of the Association for Computational Linguistics on Computational Linguistics, 1999.

[5] J. Rocchio. Relevance feedback in information retrieval. The Smart Retrieval System: Experiments in Automatic Document Processing, pages 313–323, 1971.

[6] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In Proc. of the 19th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 4–11, 1996.

[7] V. Vinay, K. Wood, N. Milic-Frayling, and I. J. Cox. Comparing relevance feedback algorithms for web search. In Proc. of the 14th Intl. Conf. on World Wide Web, 2005.

[8] M.C. Kim, K. Choi. A comparison of collocation based similarity measures in query expansion. Information Processing and Management, 35:19-30, 1999

[9] S. Gauch, J. Wang, S.M. Rachakonda. A corpus analysis approach for automatic query expansion and ist extension to multiple databases. ACM Transaction on Information Systems, 17(3):250250, 1999

[10] S.C. Wang, Y.Tanaka. Topic-oriented query expansion for web search. In Proc. Of the 15th Intl. Conf. on World Wide Web, pp. 1029-1030, 2006