Improving Automatic Semantic Tag Recommendation through Fuzzy Ontologies

Panos Alexopoulos
iSOCO
Av. del Partenon, 16-18, 1-7, 28042, Madrid, Spain
Email: palexopoulos@isoco.com

Manolis Wallace
Department of Computer Science and Technology
University of Peloponnese
End of Karaiskaki St., 22100, Tripolis, Greece
Email: wallace@uop.gr

***Abstract*—Semantic tagging of a textual document involves identifying and assigning to it appropriate entities that best summarize its content, i.e. entities that constitute a representative description of what the document is specifically about. The effective automation of this process requires from the system to be able to distinguish between the entities that play a central role to the documents's meaning and those that are just complementary to it. For example, a news article might make reference to many politicians even when its primary subject is only one of them. To that end, various approaches have utilized ontologies as a means to narrow down the meaning of a document and infer appropriate tags, including a recent contribution of ours regarding a tagging framework that exploits ontological relations. In this work we revise and extend this framework so as to be able to exploit also fuzzy ontological information. Experiments in different domains show that this exploitation manages to improve the effectiveness of the tagging process.**

I. Introduction

With the rapidly increasing popularity of Social Media sites, a lot of user-generated content has been injected in the Web, leaving a large amount of both multimedia items and textual data in the need of effective and efficient ways to organize, summarize, and search them [7]. Towards the satisfaction of that need, there is an increasing research interest in effective automatic tag recommendation techniques that may assist users in finding appropriate tags for their content, i.e. terms and phrases that are representative of the contents intended meaning [16].

The term "representative" may have a different interpretation depending on the reason why tagging is employed. As suggested in [13], tagging is typically used with the explicit intent of:

- Identifying the concepts to which various terms and phrases of the document belong (e.g. detecting terms that are locations or persons).
- Classifying a document by means of concepts that represent meaningful categories for the document user (e.g. determining whether a document is about military events or films).
- Characterizing a document by means of proper adjectives that denote some kind of judgment(e.g. "positive", "negative").
- Summarizing a document's content by means of keywords that constitute a representative description of what the document is specifically about (e.g. determining for which sport event or for which film is a document about).

The focus of this paper is on the latter intent, meaning that, for example, we do not wish just to determine whether a given document is about films or military events but identify **which specific films or events it is about**. In this kind of identification there are two important challenges:

1) To be able to distinguish between the terms that play a central role to the documents's meaning and those that are just complementary to it. For example, a news article might make reference to many politicians even when its primary subject is only one of them.
2) To be able to infer appropriate tags for the text even in the absence of explicit mentions of them within it. For example, a text describing a given military conflict is definitely relevant to the location this conflict took place, even if this location is not mentioned within it.

One way to see this problem is as a text categorization task, with each tag as a category and all documents that have the tag as training samples. Using these one can train a classifier using supervised machine learning algorithms based on various texture features extracted from them and then predict the categories of new documents using the trained classifier. Nevertheless, with each possible tag as a category, the category list becomes too large and the required training data too difficult to find and handle. Therefore, traditional supervised machine learning algorithms are not applicable for automatic tag recommendation.

With that in mind, we have already proposed in a previous work [2] a framework for automatically generating and recommending to users tags for text documents through the exploitation of domain ontologies. The latter describe the domain(s) of the texts to be tagged and their entities serve as a source of possible tags for them. The basic premise upon which the framework is based is that a given ontological entity is more likely to represent the text's meaning (and thus be an accurate tag) when there are many ontologically related to it entities in the text. These related entities can be seen as **evidence** whose quantitative and qualitative characteristics can be used to rank and suggest potential tags to the user.

To see why this premise makes sense, consider the text *"Annie Hall is a much better movie than Deconstructing Harry, mainly because Alvy Singer is such a well formed character and Diane Keaton gives the performance of er life"*. In this, the evidence provided by the entities *"Alvy Singer"* (a character in the movie Annie Hall), and *"Diane Keaton"*(an actress in the movie Annie Hall) indicates that Annie Hall is more likely to be the movie the text is about rather than Deconstructing Harry. Experimental evaluation in a large number of texts in the film domain, illustrated the effectiveness and usefulness of this approach [2].

In this paper we revise and extend the above framework so as to enable it to exploit also fuzzy ontological information, based on the assumption that the fuzziness that may characterize some of the ontology's relations can increase the evidential power of its entities and consequently the effectiveness of the tag recommendation process.

More specifically, Fuzzy Ontologies [3] are extensions of classical ontologies that, based on principles of Fuzzy Set Theory [8], allow the assignment of truth degrees to vague ontological elements in an effort to quantify their vagueness. Thus, for example, whereas in a traditional ontology one would claim that '*'Annie Hall is a comedy"* or that '*'Woody Allen is an expert director at human relations"*, in a fuzzy ontology one would claim that *"Annie Hall is a comedy to a degree of 0.7"* and that *"Woody Allen is an expert director at human relations 0.8"*.

Using a fuzzy ontology one can represent useful semantic information for the tag recommendation task in a higher level of granularity than with a crisp ontology, by taking advantage the truth degree representation capabilities of the former. For example, in the film domain, instead of having just the relation *hasPlayedInFilm*(Actor, Film) it is more useful to have the fuzzy relation *wasAnImportantActorInFilm*(Actor, Film) and relate specific actors to film using fuzzy degrees (e.g. "Robert Duvall was an important actor in Apocalypse Now to a degree of 0.6"). To see why this is the case consider the text *"Robert Duvall's brilliant performance in the film showed that his choice by Francis Ford Copola was wise"*. If Duvall and Copola have collaborated in more than one film but in only one of them Duval had a major role (as captured by the fuzzy degree of his relation to the film) then this film is more likely to be the subject of this text.

Given that, our proposed framework assumes the availability of a fuzzy ontology for the domain of the texts to be tagged and defines two components:

- A **Tag Fuzzy Ontological Evidence Model** that contains entities that may serve as tag-related evidence for the application scenario and domain at hand. Each entity is assigned evidential power degrees which denote its usefulness as evidence for the tag recommendation task.
- A **Tag Recommendation Process** that uses the evidence model to determine, for a given text, the ontological entities that potentially represent its content. A confidence score for each entity is used to denote the most probable tags.

The rest of the paper is organized as follows. In section II we present in a detailed manner the components of our proposed fuzzy tagging framework including the tag fuzzy ontological evidence model and the tag recommendation process. In section III we present and discuss experimental results regarding the the method's increased effectiveness in recommending tags when fuzziness is considered. Finally, in section IV we present related work and in section V we list our concluding remarks and outline potential future work.

II. Fuzzy Tagging Framework

A. Tag Fuzzy Ontological Evidence Model

For the purposes of this paper we define a fuzzy ontology as a tuple $O_F = \{C, R, I, i_C, i_R\}$, where

- C is a set of concepts.
- I is a set of instances.
- R is a set of fuzzy binary relations that may link pairs of concept instances.
- i_C is a concept instantiation function $C \to I$.
- i_R is a fuzzy relation instantiation function $R \times I \to [0, 1]$.

Given a fuzzy ontology, the **Tag Fuzzy Ontological Evidence Model** defines for each ontology instance that is a candidate tag, which other instances and to what extent "support" this candidacy. More formally, given a domain ontology O_F and a set of tags $T \subseteq I$, a tag fuzzy ontological evidence model is defined as a function $ftem : T \times I \to [0, 1]$. If $t \in T$ and $i \in I$ then $ftem(t, i_)$ is the degree to which the existence, within the text, of i should be considered an indication that t is a correct tag for the text.

In order to determine the above functions for a given domain and scenario we need to consider the concepts whose instances are directly or indirectly related to tags and which are expected to be present in the text to be analyzed. This means in turn that some a priori knowledge about the domain and content of the text(s) should be available. The more domain specific the texts are, the smaller the ontology needs to be and the more effective and efficient the whole resolution process is expected to be. In fact, it might be that using a larger ontology than necessary could reduce the effectiveness of the tagging process.

Thus, a strategy for selecting the minimum required instances that should be included in the tag evidence model would be the following:

- First identify the concepts whose instances may act as tag evidence in the given domain and texts.

- Then identify the subset of these concepts which constitute the central meaning of the texts and thus "determine" mostly their tag scope.
- Finally, use these concepts in order to limit the number of possible tags that may appear within the text as well as the number of instances of the other evidential concepts.

For example, let's assume that we want to tag historical texts with the conflicts they are about. In this domain and scenario, some concepts whose instances may act as evidence for conflicts are related locations, other conflicts, and persons that participated in them. The central concept in this scenario would be the military conflict, so from all the possible locations, conflicts and persons we consider only those that are related to some conflict.

In general, the result of the above process should be a tag evidence mapping function $tem : C \rightarrow R^n$ which given an evidential concept $c \in C$ returns the relations $\{r_1, r_2, ..., r_n\} \in R^n$ whose (fuzzy) composition links c's instances to tags. Table I shows such a mapping for the example of military conflicts mentioned above.

Table I
TAG EVIDENCE MAPPING FUNCTION FOR MILITARY CONFLICTS

Evidence Concept	**Tag Linking Fuzzy Relation(s)**
Military Conflict	*tookPlaceNearLocation*
Military Person	*playedMajorRoleInConflict, tookPlaceNearLocation*
Location	*isNearToLocation*

Using this mapping function, we can then calculate the tag evidence model $ftem$ as follows: Given a tag $t \in T$ and an instance $i \in I$, which belongs to some concept $c \in C$ and is related to t through the composition of the fuzzy relations $\{r_1, r_2, ..., r_n\} \in tem(c)$, we derive i) the set of instances $I_{amb} \subseteq I$ which share common identifiers with i and ii) the set of tags $T_i \subseteq T$ which are related to i through the composite relation $[r_1 \circ^t r_2 \circ^t ... \circ^t r_n]$. Then the value of the function $ftem$ for this tag and this instance is computed as follows:

$$ftem(t, i) = \frac{[r_1 \circ^t r_2 \circ^t ... \circ^t r_n](i, t)}{|I_{amb}| * \sum_{t' \in T_i} [r_1 \circ^t r_2 \circ^t ... \circ^t r_n](i, t')} \tag{1}$$

The intuition behind this formula is that the evidential power of a given instance is analogous to the fuzzy degree of its (composite) relation to the tag and inversely analogous to its own ambiguity as well as to the number and fuzzy degrees of all the tags it is is related to.

B. Tag Recommendation Process

The tag recommendation process for a given text document and a tag evidence model works as follows:

First we extract from the text the set of terms Tr that match to some $i \in I$ along with a term-meaning mapping function $m : Tr \rightarrow I$ that returns for a given term $tr \in Tr$ the instances it may refer to. We also consider I_{text} to be the superset of these instances. Then we consider as candidate tags those for which there is evidence within the text, that is all $t \in T$ for which $ftem(t, i) > 0, i \in I_{text}$. We call this set T_{cand}. Then, for a given candidate tag $t \in T_{cand}$ we compute the tag support it receives from the terms found within the text as follows:

$$sup(t, tr) = \frac{1}{|m(t)|} * \sum_{i \in m(tr)} ftem(t, i) \tag{2}$$

Finally, we compute the confidence that t is a correct tag for the text as follows:

$$c(t) = \frac{\sum_{tr \in Tr} K(t, tr)}{\sum_{t \in T_{cand}} \sum_{tr \in Tr} K(t, tr)} * \sum_{tr \in Tr} sup(t, tr) \tag{3}$$

where $K(t, tr) = 1$ if $sup(t, tr) > 0$ and 0 otherwise.

III. EXPERIMENTAL EVALUATION

To illustrate the added value that the exploitation of fuzziness brings to the tag recommendation task, we performed a set of comparative experiments in two different scenarios, the first involving texts in the film domain (film reviews) and the second texts describing military conflicts.

In the first case we focused on tagging the review texts with the film their review was actually about. Although we had available a set of 25000 IMDB reviews[1] to use as data, the need to have a comprehensive fuzzy ontology for them made us select only a small subset, consisting in the end of 100 reviews. These reviews regarded about 20 distinct films that were similar to each other in terms of genre, actors and directors and thus more difficult to distinguish between them in a given review. For these films we derived a crisp ontology from Freebase[2] and we created, in a manual fashion, a fuzzy version of it that comprised the following elements:

- Concepts: Film, Actor, Director, Character
- Relations: *wasAnImportantActorInFilm*(Actor, Film), *isFamousForDirectingFilm*(Director, Film), *wasCharacterInFilm*(Character, Film).

Using this ontology we defined the tag evidence mapping function of table II and we build a a tag fuzzy evidence evidence model for all pairs of films and evidential entities (actors, directors and characters). Then we applied the process of paragraph II-B and we determined for each review a ranked list of possible films it may refer to, using the confidence scores derived from equation 3. . Finally, we measured the effectiveness of the process by determining the number of correctly tagged texts, namely texts whose highest ranked films were the correct ones. For comparison purposes, we performed the same process using a crisp version of the

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data

[2] http://www.freebase.com

film ontology (i.e. all fuzzy degrees were equal to 1). Table III shows that the consideration of the domain's fuzziness managed to improve the tag recommendation effectiveness by 10%.

Table II
TAG EVIDENCE MAPPING FUNCTION FOR FILMS

Evidence Concept	Tag Linking Fuzzy Relation(s)
Director	*isFamousForDirectingFilm*
Actor	*wasAnImportantActorInFilm*
Character	*wasCharacterInFilm*

As a second experiment we focused on tagging a set of 100 texts describing military conflicts with the conflicts they were actually about. This time, the fuzzy ontology we created was based on DBPedia and comprised the following elements:

- Concepts: Location, Military Conflict, Military Person
- Relations: *tookPlaceNearLocation*(Military Conflict, Location), *wasAnImportantPartOfConflict*(Military Conflict, Military Conflict), *playedMajorRoleInConflict*(Military Person, Military Conflict), *isNearToLocation*(Location, Location).

The tag evidence mapping function for that scenario was that of table I and the evaluation process we followed was similar to the one about the film reviews. In this case the improvement achieved by the consideration of fuzziness was 13%.

Table III
TAG RECOMMENDATION EVALUATION RESULTS

Approach	Film Reviews	Historical Texts
No Fuzziness	80%	72%
Fuzziness	90%	85%

IV. RELATED WORK

The tagging framework presented in this paper generates tag recommendations based on fuzzy domain ontologies as a source of evidential knowledge about the correct tags. In the relevant liter there are many examples of tag recommender systems [10] [17] [5] [6] [12], but only few of them use ontologies and practically none fuzzy ones.

A work in which ontologies are used for tagging is that of [18] where the authors use a hierarchical news ontology as a common language for content based filtering in order to classify news items and to deliver personalized newspaper services on a mobile reading device. In another work [13] the authors propose a tag recommendation process based on keyphrase extraction and ontology reasoning. In particular, their approach involves the utilization of linguistic and statistical processing for determining keyphrases that could be potential tags and the exploitation of domain ontologies for suggesting tags that are not present within the document. For the latter, they use a reasoning mechanism based on the subsumption relationship between concepts (is-a) and the spreading activation algorithm of [14].

A similar approach is presented in [11] where the authors discuss ontology-based document annotation for the purpose of semantic indexing and retrieval. The method they propose expands, both syntactically and semantically, concept descriptions taken from the domain ontology in order to enhance matching in the retrieval process. The syntactic expansion is based on lexical resources (e.g. Wordnet) while the semantic one on a concept exploration algorithm that is applied on the ontology.

In [4] the authors propose GoNTogle, a framework for document annotation and retrieval, built on top of Semantic Web and Information Retrieval technologies. For the annotation part, GoNTogle supports the automatic annotation of a whole document or parts of it with ontology concepts through a learning method based on *weighted kNN* classification that exploits user annotation history and textual information to automatically suggest annotations for new documents.

In [1] the authors suggest an approach to generate semantic tag recommendations for documents based on Semantic Web ontologies and Web 2.0 services. In particular, their proposed process starts with the extraction of document entities through the utilization of Web 2.0 services (such as Yahoo's Term Extraction service and their transformation into a topic map using SKOS vocabulary (Simple Knowledge Organisation System) [9]. Then, the topics of this topic map are matched, based on document classification methods, to instances of some domain ontology expressed according to the PIMO ontology [15]. The matching pairs are shown to the users as tag recommendations and they decide whether to accept or reject them.

V. CONCLUSIONS & FUTURE WORK

In this paper we proposed a novel framework that exploits fuzzy semantic information for automatically generating and recommending semantic tags for text documents in an effort to summarize the intended meaning of their content. Our approach has been based on the customized utilization of fuzzy domain-specific ontological relations for extracting and evaluating "evidence" from within the text that may identify the correct tag(s) in the given tagging scenario.

The added value that the exploitation of fuzziness brought to the tag recommendation task was experimentally tested through experiments in different domains where the effectiveness of the method using fuzziness was measured and compared to the one without fuzziness. The results verified our intuition that through a fuzzy ontology one can represent useful semantic information for the tag recommendation task in a higher level of granularity than with a crisp ontology.

As one important obstacle for the wider applicability of our approach is the bottleneck of acquiring (through development or reuse) the required fuzzy ontological information for the domain at hand. For that reason, our future work will focus on determining automated methods for fuzzifying crisp ontological facts through various approaches, including data mining, social network analysis and crowdsourcing.

References

[1] B. Adrian, L. Sauermann, and T. Roth-Berghofer. Contag: A semantic tag recommendation system. In T. Pellegrini and S. Schaffert, editors, Proceedings of I-Semantics' 07, pages pp. 297-304. JUCS, 2007.

[2] P. Alexopoulos, J. Pavlopoulos, M. Wallace, and K. Kafentzis. 2011. Exploiting ontological relations for automatic semantic tag recommendation. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), ACM, New York, USA.

[3] Alexopoulos P., Wallace M.,Kafentzis K. and Askounis D. (2011) A Methodology for Developing Fuzzy Ontologies, Knowledge and Information Systems, pp. 1-29, Springer.

[4] N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. Sellis. Integrating keywords and semantics on document annotation and search. In Proceedings of the 2010 international conference on On the move to meaningful internet systems: Part II, OTM'10, pages 921-938, Berlin, Heidelberg, 2010. Springer-Verlag.

[5] J. Gemmell, T. Schimoler, B. Mobasher, and R. Burke. Hybrid tag recommendation for social annotation systems. In Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10, pages 829-838, New York, NY, USA, 2010. ACM.

[6] K.-n. Hassanali and V. Hatzivassiloglou. Automatic detection of tags for political blogs. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, WSA '10, pages 21-22, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[7] A. M. Kaplan and M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Business Horizons, vol. 53, no. 1, pp. 5968, 2010.

[8] Klir G, Yuan B (1995) Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall.

[9] A. Miles and D. Brickley. Skos core vocabulary specication. W3c working draft, World Wide Web Consortium, 2005.

[10] G. Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In WWW '06: Proceedings of the 15th international conference on World Wide Web, pages 953-954, New York, NY, USA, 2006. ACM Press. paper presented at the poster track.

[11] S. Nesic, F. Crestani, M. Jazayeri, and D. Gasevic. Concept-based semantic annotation, indexing and retrieval of office-like document units. In Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO '10, pages 134-135, Paris, France, 2010.

[12] B. Oliveira, P. Calado, and H. S. Pinto. Automatic tag suggestion based on resource contents. In Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns, EKAW '08, pages 255-264, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] N. Pudota, A. Dattolo, A. Baruzzo, F. Ferrara, and C. Tasso. Automatic keyphrase extraction and ontology mining for content-based tag recommendation. International Journal of Intelligent Systems, 25(12):1158-1186, 2010.

[14] M. R. Quillian. Semantic memory. In M. Minsky, editor, Semantic Information Processing, pages 227-270. MIT Press, 1968.

[15] L. Sauermann. Pimo-a pim ontology for the semantic desktop (draft). Draft, DFKI, 2006.

[16] S. C. Sood and K. Hammond, TagAssist: Automatic Tag Suggestion for Blog Posts, in International Conference on Weblogs and Social, 2007.

[17] M. Tatu, M. Srikanth, and T. D'Silva. Tag recommendations using bookmark content. In Proceedings of ECML PKDD Discovery Challenge (RSDC08), pages 96-107, 2008.

[18] L. Tenenbaum, B. Shapira, and P. Shoval. Ontology-based classication of news in an electronic newspaper. In Proceedings of INFOS 2008, Varna, Bulgaria, pages 89-97, 2008.