Sense Aware Searching and Exploration with MyTag[*]

Klaas Dellschaft
ISWeb Working Group
University Koblenz-Landau
D-56070, Koblenz, Germany
klaasd@uni-koblenz.de

Olaf Göerlitz
ISWeb Working Group
University Koblenz-Landau
D-56070, Koblenz, Germany
goerlitz@uni-koblenz.de

Martin Szomszor
School of Electronics and
Computer Science
University of Southampton
SO17 1BJ, UK
mns2@ecs.soton.ac.uk

ABSTRACT

In this work, we describe our approach on how to deal with tag ambiguity in tagging systems and how to enable a sense aware or semantic search. The sense aware search is realized by means of a Sense Repository which returns for given terms a list of potential senses. This list is then presented to the user of the cross-folksonomy search engine MyTag so that he can explicitly select the sense he wants to search for. The search results are then ranked according to this sense so that relevant resources appear higher in the result list.

1. INTRODUCTION

Nowadays Web 2.0 platforms like del.icio.us [1], Flickr [2] and YouTube [3] provide large amounts of resources such as social bookmarks, photographs and videos. Common to the platforms is the classification by so called tags that can be used for organization and retrieval. A current limitation of tagging systems is their confinement to a single media type. In previous work [4], we presented the MyTag platform[1] which allows for a personalized search and exploration in several tagging systems in parallel.

A further limitation of tagging systems is the lack of semantics which does not allow to differentiate between the senses of a tag during annotating and searching resources. In this work, we describe our approach on how to deal with the problem of tag ambiguity. In Section 2 we describe the TAGora Sense Repository which provides information about senses of a tag or term as RDF via a REST-style interface. Then, in Section 3 we describe how MyTag uses this web service for offering a sense aware search.

2. EXTRACTING WORD SENSES

The TAGora Sense Repository[2] (TSR) is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. The Sense Repository is queried by forming a REST-style URI which contains the tag (e.g. `http://tagora.ecs.soton.ac.uk/tag/apple/rdf`). The Sense Repository processes the given tag, grounds it to a set of DBPedia.org resources, and returns the results as an RDF document.

For building the Sense Repository, we processed an XML dump of all English Wikipedia pages and analysed all titles, redirection links and disambiguation pages. Additionally, for each page title we extracted and indexed a lower case version, and a concatenated version of the title (e. g. the title Second_life becomes secondlife). This style of multiple title indexing enables us to match more easily tags that are made up of compound terms. We also extracted redirection links, disambiguation links, as well as the frequencies of terms contained in the page. The results are put into a Triple Store using an extended version of the DBPedia ontology[3]. Furthermore, the data set links all Wikipedia pages to DBPedia resources via the owl:sameAs property.

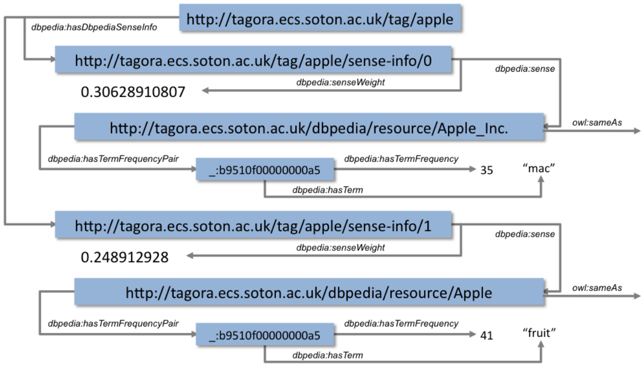

http://tagora.ecs.soton.ac.uk/tag/apple
dbpedia:hasDbpediaSenseInfo
http://tagora.ecs.soton.ac.uk/tag/apple/sense-info/0
dbpedia:sense
dbpedia:senseWeight
0.30628910807
http://tagora.ecs.soton.ac.uk/dbpedia/resource/Apple_Inc.
owl:sameAs
dbpedia:hasTermFrequencyPair
_:b9510f00000000a5
dbpedia:hasTermFrequency
35
"mac"
dbpedia:hasTerm
http://tagora.ecs.soton.ac.uk/tag/apple/sense-info/1
dbpedia:sense
dbpedia:senseWeight
0.248912928
http://tagora.ecs.soton.ac.uk/dbpedia/resource/Apple
owl:sameAs
dbpedia:hasTermFrequencyPair
_:b9510f00000000a5
dbpedia:hasTermFrequency
41
"fruit"
dbpedia:hasTerm

Figure 1: Linked data representation of tag senses

When querying the Sense Repository via the REST interface, we start by normalizing the given tag or term by removing non-alphanumeric characters, converting to lowercase characters and handling compound words [5]. Then, we query our Triple Store for a list of candidate DBPedia resources that represent possible senses of the normalized tag. During the query, we also follow the redirection and/or disambiguation links in DBPedia. Finally, a weight is associated with each possible sense. The weight is the fraction of incoming links to the Wikipedia page which is associated with that sense and the total number of incoming links for all senses of the tag. This leads to higher weights for more general senses and to lower weights for very specific senses.

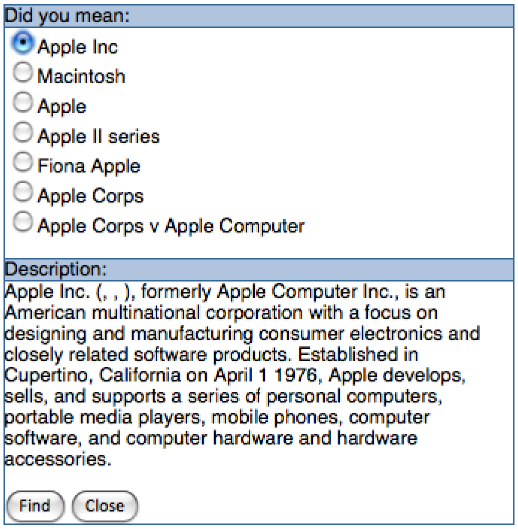
Fig. 1 provides a visual example of the linked data associated with the tag `apple` – a common tag that could refer the computer company (Apple_Inc.), or the fruit (Apple). In this example, the URI for apple (center, top) is linked to a number of sense-info instances (only two of which are shown here) via the property dbpedia:hasdbpediaSenseInfo. Each sense-info pair gives the weight (0.306 for Apple_Inc. and 0.249 for Apple) and the corresponding DBPpedia resource.

[*] This work has been supported by the European project Semiotic Dynamics in Online Social Communities (TAGora, FP6-2005-34721).

[1] `http://mytag.uni-koblenz.de/`

[2] `http://tagora.ecs.soton.ac.uk/`

[3] `http://tagora.ecs.soton.ac.uk/schemas/dbpedia`



Did you mean:

● Apple Inc
○ Macintosh
○ Apple
○ Apple II series
○ Fiona Apple
○ Apple Corps
○ Apple Corps v Apple Computer

Description:

Apple Inc. (, ,), formerly Apple Computer Inc., is an American multinational corporation with a focus on designing and manufacturing consumer electronics and closely related software products. Established in Cupertino, California on April 1 1976, Apple develops, sells, and supports a series of personal computers, portable media players, mobile phones, computer software, and computer hardware and hardware accessories.

Figure 2: MyTag sense searching dialogue

Each sense is linked to a set of blank nodes (of type termFrequencyPair) that states the frequencies of terms within the Wikipedia page of that resource.

3. SENSE-AWARE SEARCH IN MYTAG

When a user submits a query to MyTag, it not only queries the different tagging systems for relevant resources but it also queries the Tagora Sense Repository (see above). The Sense Repository returns possible senses of the search terms. The different possible senses are shown to the user if he clicks on the question mark that is appended to the search term (see Fig. 2). For each possible sense, the user sees the page title of the associated DBPedia page and the first three sentences of the English description. The user can then select the intended sense of the search term and re-rank the current list of results so that resources corresponding to the intended sense are ranked higher. A more detailed description of the complete process can be found below.

3.1 Removing Irrelevant Senses

Usually, the Sense Repository returns more possible senses for a search term than can be found in documents returned by the different tagging systems. Thus, in a first step all senses are removed from the list which are not contained in the set of documents anyway. This helps to significantly reduce the number of possible senses that are shown by the user interface. For removing irrelevant senses, we compare the *term frequencies* of each sense with the *tag cloud* of the current search results. The tag cloud contains all tags and how often they are assigned to the documents in the current search results. We remove all senses which do not have at least one tag and/or term in common between their term frequencies returned from the Sense Repository and the tag cloud of the search results.

For example, when searching for `apple` the Tagora Sense Repository returns *Gwyneth Paltrow*[4] as a possible sense because the first name of her daughter is *Apple*. Related terms for this sense are for example *daughter*, *actress* or *paltrow*. But because none of these related terms is contained in the tag cloud of the search results, this sense is discarded and not shown to the user (see Fig. 2).

3.2 Ranking Search Results

When the user selects one of the offered senses and clicks on the *Search* button in the search interface (see Fig. 2), MyTag calculates a new ranking for the documents in the result sets retrieved from the different tagging systems. For this purpose, we reuse the ranking algorithm that is also used for providing a personalized ranking of search results (see [4]). It basically calculates the cosine similarity $\mathbf{r}$ between the term frequencies vector $\mathbf{p}$ of the selected sense and the tag vector $\mathbf{v}$ of a resource:

$$\mathbf{r} = \frac{\mathbf{v} \cdot \mathbf{p}}{\|\mathbf{v}\| \cdot \|\mathbf{p}\|} \qquad (1)$$

All documents are then reordered based on their individual $\mathbf{r}$ value.

4. CONCLUSIONS AND FUTURE WORK

In this work, we presented the TAGora Sense Repository and MyTag. The Sense Repository provides a list of possible senses for a tag or term. For each sense it provides a weight which indicates whether it is a general or more specific sense of the term. Furthermore, it provides frequencies of other terms which are related to this sense. The information from the Sense Repository is used by MyTag to offer its users a list of possible senses for a search query and for ranking the resources in the search result list according to their similarity with the selected sense.

In the future, we want to extend the current approach so that not only irrelevant senses are excluded from the sense proposals but also very similar meanings and/or synonyms are merged (e. g. *Apple Inc*, *Macintosh* and *Apple II series* in Fig. 2). So far, we only looked manually at selected examples to judge whether the output and results of the disambiguation service in MyTag makes sense. These results look promising. Nevertheless, as future work, we plan a detailed evaluation of the approach. We are interested in answering the following questions: (1) Which disambiguation strategies are applied by users of search engines (e. g. adding further search terms) and how successful are they compared to our approach? (2) How often does the TAGora Sense Repository provide appropriate senses that can be used for disambiguation? (3) To which extend is the ranking algorithm able to rank relevant resources higher in the list of results?

5. REFERENCES

[1] Del.icio.us Website. `http://del.icio.us`.

[2] Flickr Website. `http://flickr.com`.

[3] YouTube Website. `http://youtube.com`.

[4] M. Braun, K. Dellschaft, T. Franz, D. Hering, P. Jungen, H. Metzler, E. Müller, A. Rostilov, and C. Saathoff. Personalized Search and Exploration with MyTag. In *Proc. of WWW 2008 Poster Session*, 2008.

[5] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In *Proc. of 7th International Semantic Web Conference*, 2008.

[4] `http://dbpedia.org/resource/Gwyneth_Paltrow`