Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations

Iván Cantador[1], Martin Szomszor[2], Harith Alani[2],
Miriam Fernández[1], Pablo Castells[1]

[1] Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
{ivan.cantador, miriam.fernandez, pablo.castells}@uam.es

[2] School of Electronics and Computer Science
University of Southampton
SO17 1BJ Southampton, United Kingdom
{mns2, ha}@ ecs.soton.ac.uk

Abstract. Many advanced recommendation frameworks employ ontologies of various complexities to model individuals and items, providing a mechanism for the expression of user interests and the representation of item attributes. As a result, complex matching techniques can be applied to support individuals in the discovery of items according to explicit and implicit user preferences. Recently, the rapid adoption of Web2.0, and the proliferation of social networking sites, has resulted in more and more users providing an increasing amount of information about themselves that could be exploited for recommendation purposes. However, the unification of personal information with ontologies using the contemporary knowledge representation methods often associated with Web2.0 applications, such as community tagging, is a non-trivial task. In this paper, we propose a method for the unification of tags with ontologies by grounding tags to a shared representation in the form of Wordnet and Wikipedia. We incorporate individuals' tagging history into their ontological profiles by matching tags with ontology concepts. This approach is preliminary evaluated by extending an existing news recommendation system with user tagging histories harvested from popular social networking sites.

Keywords: social tagging, web 2.0, ontology, semantic web, user modelling, recommender systems.

1 Introduction

The increasing proliferation of Web2.0 style sharing platforms, coupled with the rapid development of novel ways to exploit them, is paving the way for new paradigms in Web usage. Virtual communities and on-line services such as social networking, folksonomies, blogs, and wikis, are fostering an increase in user participation, engaging users and encouraging them to share more and more information, resources, and opinions. The huge amount of information resulting from this emerging phenomenon gives rise to excellent opportunities to investigate, understand, and exploit the knowledge about the users' interests, preferences and needs. However, the current infrastructure of the Web does not provide the mechanisms necessary to

consolidate this wealth of personal data since they are spread over many unconnected, heterogeneous sources.

Community tagging sites, and their respective folksonomies, are a clear example of this situation: users have access to a plethora of web sites that allow them to annotate and share many types of resources. For example, they can organise and make photos available on Flickr[1], classify and share bookmarks using del.icio.us[2], communicate and share resources with friends using Facebook[3]. Through personal tags, users implicitly declare different facets of their personalities, such as their favourite book subjects on LibraryThing[4], movie preferences on IMDb[5], music tastes on Last.fm[6], and so forth. Therefore, the domains covered by social tagging applications are both disparate and divergent, creating considerably complex and extensive descriptions of user profiles.

In the current Web2.0 landscape, there is a distinct lack of tools to support users with meaningful ways to query and retrieve resources spread over disparate end-points: users should be able to search consistently across a broad range of sites for diverse media types such as articles, reviews, videos, and photos. Furthermore, such sites could be used to support the recommendation of new resources belonging to multiple domains based on tags from different sites. As a step towards making this vision a reality, we explore the use of syntactic and semantic based technologies for the combination, communication and exploitation of information from different social systems.

In this paper, we present an approach for the consolidation of social tagging information from multiple sources into ontologies that describe the domains of interest covered by the tags. Ontology-based user profiles enable rich comparisons of user interests against semantic annotations of resources, in order to make personal recommendations. This principle has already been tested by the authors in different personalised information retrieval frameworks, such as semantic query-based searching [4], personalised context-aware content retrieval [13], group-oriented profiling [3], and multi-facet hybrid recommendations [2].

We propose to feed the previous strategies with user profiles built from personal tag clouds obtained from Flickr and del.icio.us web sites. The mapping of those social tags to our ontological structures involve three steps: the filtering of tags, the acquisition of semantic information from the Web to map the remaining tags into a common vocabulary, and the categorisation of the obtained concepts according to the existing ontology classes.

An application of the above techniques has been tested in News@hand, a news recommender system which integrates our different ontology-based recommendation approaches. In this system, ontological knowledge bases and user profiles are generated from public social tagging information, using the aforementioned techniques. The News@hand system, along with the automatic acquisition of news articles from the Web, and the automatic semantic annotation of these items using Natural Language Processing tools [1] and the Lucene[7] indexer shall also be described.

---

[1] Flickr, Photo Sharing, http://www.flickr.com/
[2] del.icio.us, Social Bookmark manager, http://del.icio.us/
[3] Facebook, Social Networking, http://www.facebook.com/
[4] LibraryThing, Personal Online Book Catalogues, http://www.librarything.com/
[5] IMDb, Internet Movie Database, http://imdb.com/
[6] Last.fm, The Social Music Revolution, http://www.last.fm/
[7] Lucene, An Open Source Information Retrieval Library, http://lucene.apache.org/

The structure of the paper is the following. Section 2 briefly describes our approach for representing user preferences and item features using ontology-based knowledge structures, and how they are exploited by several recommendation models. Section 3 explains mechanisms to automatically relate and transform social tagging and external semantic information into our ontological knowledge structures. A real implementation and evaluation of the previous tag transformation and recommendation processes within a news recommender system are presented in section 4. Finally, section 5 proclaims some conclusions and future research lines.

2 Hybrid recommendations

In this section, we summarise the ontology-based knowledge representation and recommendation models in which filtered social tags are proposed to be integrated and exploited.

2.1 Ontology-based representation of item features and user preferences

In the knowledge representation we propose [4, 13], user preferences are described as vectors $\mathbf{u}_m = (u_{m,1}, u_{m,2}, ..., u_{m,K})$ where $u_{m,k} \in [0,1]$ measures the intensity of the interest of user $u_m \in \mathcal{U}$ for concept $c_k \in \mathcal{O}$ (a class or an instance) in a domain ontology $\mathcal{O}$, K being the total number of concepts in the ontology. Similarly, items $d_n \in \mathcal{D}$ are assumed to be annotated by vectors $\mathbf{d_n} = (d_{n,1}, d_{n,2}, ..., d_{n,K})$ of concept weights, in the same vector-space as user preferences.

The main advantages of this knowledge representation are its portability, thanks to the XML-based Semantic Web standards, the domain independency of the subsequent content retrieval and recommendation algorithms, and the multi-source nature of the proposal (different types of media could be annotated: texts, images, videos).

2.2 Personalised content retrieval

Our notion of content retrieval is based on a matching algorithm that provides a personal relevance measure $pref(d_n, u_m)$ of an item d_n for a user u_m. This measure is set according to semantic preferences of the user and semantic annotations of the item, and is based on a cosine vector similarity $\cos(\mathbf{d}_n, \mathbf{u}_m)$. The obtained similarity values (Personalised Ranking module of Figure 1) can be combined with query-based scores without personalisation $sim(d_n, q)$ and semantic context information (Item Retrieving module of Figure 1), to produce combined rankings [13].

To overcome the existence of *sparsity* in user profiles, we propose a preference spreading mechanism, which expands the initial set of preferences stored in user profiles through explicit semantic relations with other concepts in the ontology. Our approach is based on Constrained Spreading Activation (CSA), and is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed. We have empirically demonstrated [3, 13] that preference extension improves retrieval precision and recall. It also helps to mitigate other well-known limitations of recommender systems such as the cold-start, overspecialisation and portfolio effects.

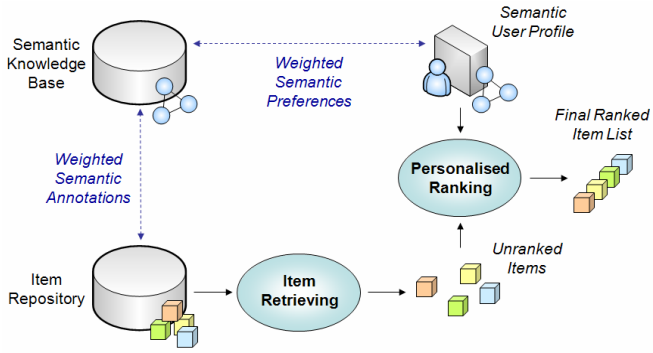

Figure 1. Ontology-based personalised content retrieval

2.3 Context-aware recommendations

The context is represented in our approach [13] as a set of weighted ontology concepts. This set is obtained by collecting the concepts that have been involved in the interaction of the user (e.g. accessed items) during a session. It is built in such a way that the importance of concepts fades away with time by a decay factor. Once the context is built, a contextual activation of user preferences is achieved by finding semantic paths linking preferences to context. These paths are made of existing relations between concepts in the ontologies, following the spreading technique mentioned in section 2.2.

2.4 Group-oriented recommendations

The presented user profile representation allows us to easily model groups of users. We have explored the combination of the ontology-based profiles to meet this purpose [3], on a per concept basis, following different strategies from social choice theory. In our approach, user profiles are merged to form a shared group profile, so that common content recommendations are generated according to this new profile.

2.5 Multi-facet hybrid recommendations

In order to make hybrid recommendations we cluster the semantic space based on the correlation of concepts appearing in the profiles of individual users. The obtained clusters C_q represent groups of preferences (topics of interests) shared by a significant number of users. Using these clusters profiles are partitioned into semantic segments. Each of these segments corresponds to a cluster and represents a subset of the user interests that is shared by the users who contributed to the clustering process. By thus introducing further structure in user profiles, we define relations among users at different levels, obtaining multilayered communities of interest.

Exploiting the relations of the communities which emerge from the users' interests, and combining them with item semantic information, we have presented in [2] several recommendation models that compare the current user interests with those of the others users in a double way. First, according to item characteristics, and second, according to connections among user interests, in both cases at different semantic layers.

$$pref\left(d_n,u_m\right)=\sum_q nsim\left(d_n,C_q\right)\sum_i nsim_q\left(u_m,u_i\right)\cdot sim_q\left(d_n,u_i\right)$$

3 Relating social tags to ontological information

Parallel to the proliferation and growth of social tagging systems, the research community is increasing its efforts to analyse the complex dynamics underlying folksonomies, and investigate the exploitation of this phenomenon in multiple domains. Results reported in [5] suggest that users of social systems share behaviours which appear to follow simple tagging activity patterns. Understanding, predicting and controlling the semiotic dynamics of online social systems are the base pillars for a wide variety of applications.

For these purposes, the establishment of a common vocabulary (set of tags) shared by users in different social systems is a desirable situation. Indeed, recent works have focused on the improvement of tagging functionalities to generate tag datasets in a controlled, coordinated way. P-TAG [6] is a method that automatically generates personalised tags for web pages, producing keywords relevant both to their textual content and to data collected from the user's browsing. In [8], an adaptation of user-based collaborative filtering and a graph-based recommender is presented as a tag recommendation mechanism that eases the process of finding good tags for a resource, and consolidating the creation of a consistent tag vocabulary across users.

The integration of folksonomies and the Semantic Web has been envisioned as an alternative approach to the collaborative organisation of shared tagging information. The proposal presented in [11] uses a combination of pre-processing strategies and statistical techniques together with knowledge provided by ontologies for making explicit the semantics behind the tag space in social tagging systems.

In the work presented herein, we propose the use of knowledge structures defined by multiple domain ontologies as a common semantic layer to unify and classify social tags from several Web 2.0 sites. More specifically, we propose a mechanism for the creation of ontology instances for the gathered tags, according to semantic information collected from the Web. Tagging information is linked to ontological structures by our method through a sequence comprising three processing steps:

- *Filtering social tags*: To facilitate the integration of information from different social sources as well as the subsequent translation of that information into ontological knowledge, a pre-processing of the tags is needed, associating them to a common vocabulary, shared by the different involved applications. Morphologic and semantic transformations of tags are performed at this stage based on the WordNet English dictionary [9], the Wikipedia[8] encyclopaedia and the Google[9] web search engine.
- *Obtaining semantic information about social tags*: The shared vocabulary is created with the use of Wikipedia, which provides semantic information about millions of concepts.
- *Categorisation of social tags into ontology classes*: Once the tags have been filtered and mapped to a shared vocabulary, they are automatically converted into instances of classes of domain ontologies. Again, semantic categorisation information available in Wikipedia is exploited in this process.

These steps are explained in more detail in the next subsections.

[8] Wikipedia, The Free Encyclopaedia, http://en.wikipedia.org/
[9] Google, Web Search Engine, http://www.google.com/

3.1 Filtering social tags

Raw tagging information can be noisy and inconsistent. When manual tags are introduced with a non-controlled tagging mechanism, people often make grammatical mistakes (e.g. *barclona* instead of *barcelona*), tag concepts indistinctly in singular, plural or derived forms (*blog*, *blogs*, *blogging*), sometimes add adjectives, adverbs, prepositions or pronouns to the main concept of the tag (*beautiful car*, *to read*), or use synonyms and acronyms that could be converted into a single tag (*biscuit* and *cookie*, *ny* and *new york*). Moreover, the tag encoding and storage mechanisms used by social systems often alter the tags introduced by the users: they may transform white spaces (*san francisco*, *san-francisco*, *san_francisco*, *sanfrancisco*) and special characters in the tags (*los angeles* for *los ángeles*, *zurich* instead of *zürich*), etc.

Thus, while it is possible to gather information from multiple folksonomy sites, such as Flickr or del.icio.us, inconsistency will lead to confusion and loss of information when tagging data is compared. For example, if a user has tagged photos from a recent holiday in New York with nyc, but also bookmarked relevant pages in del.icio.us with new_york, the correlation will be lost. In order to facilitate the folksonomy data analysis and integration, tags have to be filtered and mapped to a shared vocabulary. Here, we present a tag filtering architecture that makes use of external knowledge resources such as the WordNet dictionary, Wikipedia encyclopaedia and Google web search engine.

The filtering process is a sequential execution where the output from one filtering step is used as input to the next. The output of the entire filtering process is a set of new tags that correspond to an agreed representation. As will be explained below, this is achieved by correlating tags to entries in two large knowledge resources: Wordnet and Wikipedia. Wordnet is a lexical database and thesaurus that group English words into sets of cognitive synonyms called synsets, providing definitions of terms, and modelling various semantic relations between concepts: synonym, hypernym, hyponym, among others. Wikipedia is a multilingual, open-access, free-content encyclopaedia on the Internet. Using a wiki style of collaborative content writing, is has grown to become one of the largest reference Web sites with over 75,000 active contributors, maintaining approximately 9,000,000 articles in over 250 languages (as of February 2008). Wikipedia contains collaboratively generated categories that classify and relate entries, and also supports term disambiguation and dereferencing of acronyms.

Figure 2 provides a visual representation of the filtering process where a set of raw tags are transformed into a set of filtered tags and a set of discarded tags. Each of the numbers in the diagram corresponds to a step outlined below.

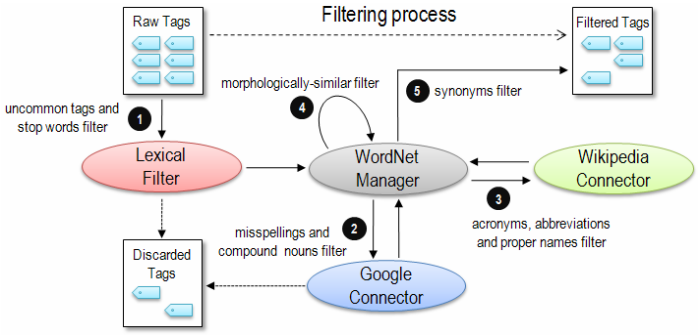

Figure 2. The tag filtering process

For this work, tags from public available user accounts from Flickr and del.icio.us sites have been collected and filtered. A total of 1004 user profiles have been gathered from these two systems, providing 149,529 and 84,851 distinct tags respectively. Initially, the intersection between both datasets was 28,550 common tags.

Step 1: Lexical filtering

After raw tags have been harvested from different folksonomy sites, they are passed to the *Lexical Filter*, which applies several filtering operations. Tags that are too small (with length = 1) or too large (length > 25) are removed, resulting in a discarding rate of approximately 3% of the initial dataset. In addition, considering the discrepancies in the use of special characters (such as accents, dieresis and caret symbol), we convert such special characters to a base form (e.g., the characters à, á, â, ã, ä, å are converted to a).

Tags containing numbers are also filtered based on a set of custom heuristics. For example, to maintain salient numbers, such as dates (*2006*, *2007*, etc), common references (*911*, *360*, *666*, etc), or combinations of alphanumeric characters (*7 up*, *4 x 4*, *35 mm*), we discard unpopular tags below a certain global tag frequency threshold. Finally, common stop-words, such as pronouns, articles, prepositions and conjunctions are removed. After lexical filtering, tags are passed on to the *Wordnet Manager*. If a tag has an exact match in Wordnet, we pass it on directly to the set of filtered tags, to save further unnecessary processing.

Step 2: Compound nouns and misspellings

If a tag is not found in Wordnet, we consider possible misspellings and compound nouns. Motivated by [11], to solve these problems, we make use of the Google "did you mean" mechanism. When a search term is entered, the Google engine checks whether more relevant search results are found with an alternative spelling. Because Google's spell check is based on occurrences of all words on the Internet, it is able to suggest common spellings for proper nouns that would not appear in a standard dictionary.

The Google "did you mean" mechanism also provides an excellent way to resolve compound nouns. Since most tagging systems prevent users from entering white spaces into the tag value, users create compound nouns by concatenating nouns together or delimiting them with a non-alphanumeric character such as _ or -, which introduces an obvious source of complication when aligning folksonomies. By sending compound nouns to Google, we easily resolve the tag into its constituent parts. This mechanism works well for compound nouns with two terms, but is likely to fail if more than two terms are used. For example, the tag *sanfrancisco* is corrected to *san francisco*, but the tag *unitedkingdomsouthampton* is not resolved by Google.

We have thus developed a complementary algorithm that quickly and accurately splits compound nouns of three or more terms. The main idea is to firstly sort the tags in alphabetical order, and secondly process the generated tag list sequentially. By caching previous lookups, and matching the first shared characters of the current tag string, we are able to split it into a prefix (previously resolved by Google) and a postfix. A second lookup is then made using the postfix to seek further possible matches. The process is iteratively repeated until no splits are obtained from our *Google Connector*. Compared to a bespoke string-splitting heuristic, this process has a very low computational cost. This mechanism successfully recognizes long compound

nouns such as *war of the worlds*, *lord of the rings*, and *martin luther king jr*.

Similarly to Step 1, after using Google to check for misspellings and compound nouns, the results are validated against the *Wordnet Manager*. Unprocessed tags are added to the pending tag stack, and unmatched tags are discarded.

Step 3: Wikipedia correlation

Many of the popular tags occurring in community tagging systems do not appear in grammar dictionaries, such as Wordnet, because they correspond to proper names (such as famous people, places, or companies), contemporary terminology (such as *web2.0* and *podcast*), or are widely used acronyms (such as *asap* and *diy*).

In order to provide an agreed representation for such tags, we correlate tags to their appropriate Wikipedia entries. For example, when searching the tag *nyc* in Wikipedia, the entry for New York City is returned. The advantage of using Wikipedia to agree on tags from folksonomies is that Wikipedia is a community-driven knowledge base, much like folksonomies are, so that it rapidly adapts to accommodate new terminology.

Apart from consolidating agreed terms for the filtered tags, our *Wikipedia Connector* retrieves semantic information about each obtained entry. Specifically, it extracts ambiguous concepts (e.g., "java programming language" and "java island" for the entry "java"), and collaboratively generated categories (e.g., "living people", "film actors" and "american male models" for the entry "brad pitt"). This information is exploited by the ontology population and annotation processes described below.

Step 4: Morphologically similar terms

An additional issue to be considered during the filtering process is that users often use morphologically similar terms to refer to the same concept. One very common example of this is the no discrepancy between singular and plural terms, such as *blog* and *blogs*, and other morphological deviations (e.g. *blogging*). In this step, using a custom singularisation algorithm, and the stemming functions provided by the Snowball library[10], we reduce morphologically similar tags to a single tag. For each group of similar tags, the shortest term found in Wordnet is used as the representative tag.

Step 5: WordNet synonyms

When people communicate a certain concept, they often use synonyms, i.e., terms that have the same meaning, but with different morphological forms. A natural filtering step is the simplification of the tag sets by merging pairs of synonyms into single terms.

WordNet provides synonym relations between synsets of the terms. However, due to ambiguous meanings of the tags, not all of them can be taken into consideration, and the filtering process must be very carefully executed. Our merging process comprises three stages. In the first stage, a matrix of synonym relations is created by using Wordnet. In the second stage, according to the number of synonym relations found for each tag, we identify the non-ambiguous synonym pairs, and finally, stage three replaces each of the synonym pairs by the term that is most popular. Examples of thus processed synonym pairs are *android* and *humanoid*, *thesis* and *dissertation*, *funicular* and *cable railway*, *stein* and *beer mug*, or *poinsettia* and *christmas flower*.

[10] Snowball, String-handling Language, http://snowball.tartarus.org/

3.2 Obtaining semantic information about social tags

In order to populate ontologies with concepts associated to the filtered social tags, general multi-domain semantic knowledge is needed. In this work, as mentioned before, we propose to extract that information from Wikipedia. The Wikipedia articles describe a number of different types of entities: people, places, companies, etc., providing descriptions, references, and even images about the described entities.

Many of these entities are ambiguous, having several meanings for different contexts. For instance, the same tag "java" could be assigned to a Flickr picture of the Pacific island, or a del.icio.us page about the programming language. One approach to address tag disambiguation is by using the information available in Wikipedia. A Wikipedia article is fairly structured: the title of the page is the entity name itself (as found in Wikipedia), the content is divided into well delimited sections, and a first paragraph is dedicated to possible disambiguation options for the corresponding term. For example, the page of the entry "apple" starts as follows:

- "This article is about the *fruit*…"
- "For the *Beatles multimedia corporation*, see…"
- "For *the technology company*, see…"

Apart from these elements, every article contains a set of collaboratively generated categories. Hence, for example, the categories created for the concept "teide" are: world heritage sites in spain, tenerife, mountains of spain, volcanoes of spain, national parks of spain, stratovolcanoes, hotspot volcanoes, and decade volcanoes. Processing somehow the previous information, we might infer that "teide" is a volcano in Spain.

Disambiguation and categorisation information have been therefore extracted from Wikipedia for every concept appearing in our social tag datasets. Once the most suitable category for a term is determined, we match its relevant categories to classes defined in the domain ontologies, as explained next.

3.3 Categorisation of social tags into ontology classes

The assignment of an ontology class to a Wikipedia entry is based on a morphologic matching between the name and the categories of the entry, and the names of the ontology classes. The ontology classes with most similar names to the name and categories of the entry are chosen as the classes whereof the corresponding individual (instance) is to be created. The created instances are assigned a URI containing the entry name, and are given RDFS labels with the Wikipedia categories.

To better explain the proposed matching method, let us consider the following example. Let "brad pitt" be the concept we wish to instantiate. If we look up this concept in Wikipedia, a page with information about the actor is returned. At the end of the page, several categories are shown: "action film actors", "american film actors", "american television actors", "best supporting actor golden globe (film)", "living people", "missouri actors", "oklahoma (state) actors", "american male models", etc.

After retrieving that information, all the terms (tokens) that appear in the name and categories of the entry (which we will henceforth refer to as entry terms) are morphologically compared with the names of the ontology classes (assuming that a class-label mapping is available, as it is usually the case). Computing the Levenshtein distance, and applying singularisation and stemming mechanisms, only the entry terms that match

some class name, above a certain distance threshold, are kept, and the rest are discarded. For instance, suppose that "action", "actor", "film", "people", and "television" are the ones sufficiently close to some ontology class. To select the most appropriate ontology class among the matching ones, we firstly create a vector whose coordinates correspond to the filtered entry terms, taking as value the number of times the term appears in the entry name and categories together. In the example, the vector might be as follows: {(action, 1), (actor, 6), (film, 3), (people, 1), (television, 1)}, assuming that "actor" appears in six categories of the Wikipedia entry "brad pitt", and so forth.

Once this vector has been created, one or more ontology classes are selected by the following heuristic:

1. If a single coordinate holds the maximum value in the vector, we select the ontology class that matches the corresponding term.
2. In case of a tie between several coordinates having the maximum value, a new vector is created, containing the matched classes plus their taxonomic ancestor classes in the ontologies. Then the weight of each component is computed as the number of times the corresponding class is found in this step. Finally, the original classes that have the highliest valued ancestor in the new vector are selected.

Here "ontology class" and "ancestor" denote a loose notion admitting a broad range of taxonomic constructs, ranging from informally built subject hierarchies (such as the ones defined in the Open Directory tree or, in our experiments, the IPTC Subjects), to pure ontology classes in a strict Description Logic sense.

In our example, the weight for the term "actor" is the highest, so we select its matching class as the category of the entry. Thus, assuming that the class matching this term was "Actor", we finally define "Brad Pitt" as an instance of "Actor".

Now suppose that, instead, the vector for Brad Pitt was {(actor, 1), (film, 1), (people, 1)}. In that case, there would be a tie in the matching classes, and we would apply the second case of the heuristic. We take the ancestor classes, which could be e.g. "cinema industry" for "actor", "cinema industry" for "film", and "mammal" for "person", and create a weighted list with the original and ancestor classes. Then we count the number of times each class appears in the previous list, and create the new vector: {(actor, 1), (film, 1), (person, 1), (cinema industry, 2), (mammal, 1)}. Since the class "cinema industry" has the highest weight, we finally select its sub-classes "actor" and "film" as the classes of the instance "brad pitt".

We must note that our ontology population mechanism does not necessarily generate individuals following a strict semantic "is-a" schema, but a more relaxed semantic "is-related-to" association principle. This is not a problem for our final purposes in personalised content retrieval, since the annotation and recommendation methods in that area are themselves rooted on models of inherently approximated nature, e.g. regarding the relationships between concepts and item contents.

4 Preliminary evaluations

Recent works show an increasing interest in using social tagging information to enhance personalised content retrieval and recommendation. FolkRank [7] is a search algorithm that exploits the structure of folksonomies to find communities and organise search results. The recommender system presented in [10] suggests web pages available on the Internet, by using folksonomy and social bookmarking information. The movie

recommender proposed in [12] is built on keywords assigned to movies via collaborative tagging, and demonstrates the feasibility of making accurate recommendations based on the similarity of item keywords to those of the user's rating tag-clouds.

In the following, we present and preliminary evaluate how our ontological knowledge representation, recommendation models, and tag filtering and matching strategies are integrated in News@hand, a news recommender system.

4.1 News@hand

News@hand is a news recommender system that describes news contents and user preferences with a controlled and structured vocabulary, using semantic-based technologies, and integrating the recommendation models described in section 2. Figure 3 depicts how ontology-based item descriptions and user profiles are created and exploited by the system.

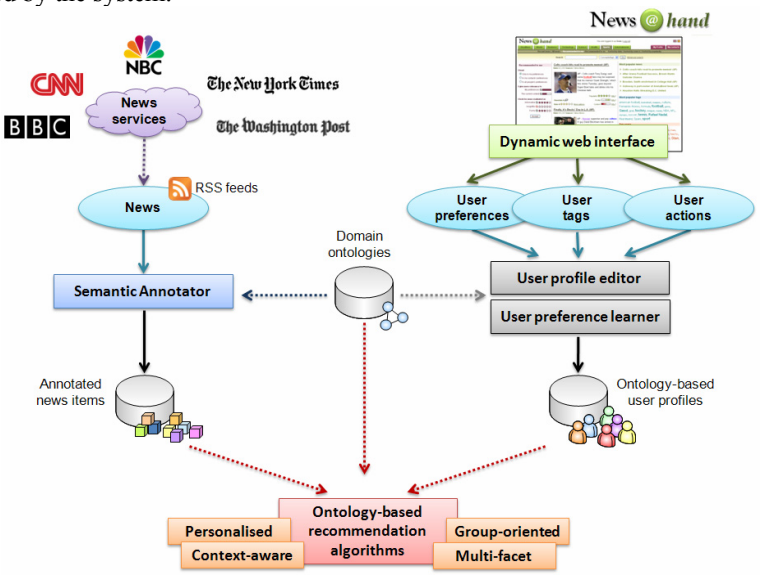

Figure 3. Architecture of News@hand

News items are automatically and periodically retrieved from several on-line news services via RSS feeds. The title and summary of the retrieved news are annotated with concepts of the domain ontologies. A dynamic graphic interface allows the system to automatically retrieve all the users' inputs in order to analyse their behaviour with the system, update their preferences, and adjust the recommendations in real time.

Figure 4 shows a screenshot of a typical news recommendation page in News@hand. The news items are classified into eight different sections: headlines, world, business, technology, science, health, sports and entertainment. When the user is not logged in the system, s/he can browse any of the previous sections, but the items are listed without any personalised criterion. On the other hand, when the user is logged in the system, recommendation and profile edition functionalities are activated, and the user can browse the news according to his and others' preferences in different ways. Click history is used to detect the short term user interests, which represent the dynamic semantic context exploited by our personalised content retrieval mechanism.

The terms occurring in the title and summary that are associated to semantic annotations of the contents, the user profile, and the current context are highlighted with different colours. A collaborative rating is shown on a 0 to 5 star scale, and two coloured bars indicate the relevance of the item for the profile and the context. The user has the possibility adding comments, tags and ratings to the article. S/he also can set parameters for single or group-oriented recommendations, such as the activation or deactivation of his/her individual preferences, those of his/her contacts and/or all other users, the weight that the dynamic context should have over the profile, and the weight of multiple rating criteria.

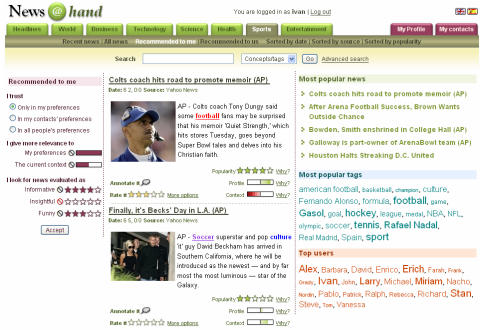

Figure 4. Item recommendation page of News@hand

4.2 Knowledge base

A total of 17 ontologies have been used for the current version of the system. They are adaptations of the IPTC ontology[11], which contains concepts of multiple domains such as education, culture, politics, religion, science, technology, business, health, entertainment, sports, weather, etc. They have been populated with semantic information about the tags we extracted from Flickr and del.icio.us web sites, applying the population mechanism explained in Section 3. A total of 137,254 Wikipedia entries were used to populate 744 ontology classes with 121,135 instances. Table 1 describes the characteristics of the obtained knowledge base.

In order to evaluate the ontology population process, we asked 20 users to randomly select, and manually assess 25 instances of each ontology. They were undergraduate and PhD students of our department, half of them with experience on ontological engineering. They were requested to declare whether each instance was assigned to its correct class, to a less correct class but belonging to a suitable ontology, or to an incorrect class/ontology. The table shows the average accuracy values for all the users considering correct class and correct ontology assignments.

These preliminary results demonstrate the feasibility of our ontology population mechanism. The average accuracy for class assignment is 69.9%, and the average accuracy for ontology assignment arises to 84.4%. Improvements in our mapping heuristics can be investigated. Nevertheless, we presume they are good enough for our recommendation goals. In general, the main common concepts are correctly instantiated, and the effect of an isolated incorrect annotation in a news item is mitigated by the domain/s of the rest of the correct annotations.

[11] IPTC ontology, http://nets.ii.uam.es/mesh/news-at-hand/news-at-hand_iptc-kb_v01.zip

Table 1. Number of classes and instances of News@hand KB, and average population accuracy

Ontology	#classes	#instances	Avg. #instances/class	Avg. accuracy
arts, culture, entertainment	87	33,278	383	78.7 / 93.3
crime, law, justice	22	971	44	62.7 / 73.3
disasters, accidents	16	287	18	74.7 / 84.0
economy, business, finance	161	25,345	157	69.3 / 80.0
education	20	3,542	177	57.5 / 76.7
environmental issues	41	20,581	502	72.0 / 85.3
health	26	1,078	41	65.3 / 89.3
human interests	6	576	96	64.0 / 84.0
labour	6	133	22	70.7 / 78.7
lifestyle, leisure	29	4,895	169	72.0 / 90.7
politics	54	3,206	59	60.0 / 81.3
religion, belief	31	3,248	105	84.0 / 90.7
science, technology	50	7,869	157	68.0 / 86.7
social issues	39	8,673	222	70.7 / 85.3
sports	124	5,567	45	72.0 / 86.7
unrests, conflicts, wars	23	1,820	79	61.3 / 80.0
weather	9	66	7	69.7 / 89.5
	744	121,135	163 (avg.)	**69.9** / **84.4**

4.3 Semantic annotation of news

News@hand periodically retrieves news items from the websites of well-known news media, such as BBC, CNN, NBC, The New York Times, and The Washington Post. These items are obtained via RSS feeds, and contain information of published news articles: their title, summary of content, publication date, hyperlinks to the full texts and related on-line images. The system analyses and automatically annotates the textual information (title and summary) of the RSS feeds (Figure 5).

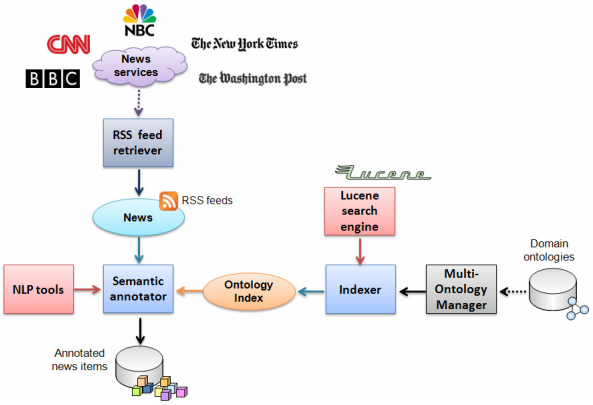

Figure 5. Automatic RSS feeds extraction and semantic annotation processes in News@hand

Using a set of Natural Language Processing tools [1], an annotation module removes stop words and extracts relevant (simple and compound) terms, categorised according to their Part of Speech (PoS): nouns, verbs, adjectives or adverbs. Then, nouns are morphologically compared with the names of the classes and instances of the domain ontologies. The comparisons are done using an ontology index created with Lucene, and

according to fuzzy metrics based on the Levenshtein distance. For each term, if similarities above a certain threshold are found, the most similar semantic concept (class or instance) is chosen and added as an annotation of the news item. After all the annotations are created, a TF-IDF based technique computes and assigns them weights.

For 2 months, since 1st January 2008, we have been daily gathering RSS feeds. A total of 9,698 news items were stored. For this dataset, we run our semantic annotation mechanism, and a total of 66,378 annotations were obtained. Table 2 shows a summary of the average number of annotations per news item generated with our system. Similarly to the experiments conducted for our ontology population strategy, we asked the 20 students to evaluate 5 news items from each of the 8 topic sections of News@hand, giving ratings with values from 0 to 10. The annotation accuracies for each topic are also presented in the table. An average accuracy of 74.8% was obtained.

Table 2. Average number of annotations per news item, and average annotation accuracies

	headlines	**world**	**business**	**technology**	**science**	**health**	**sports**	**entertainment**
#news items	2,660	2,200	1,739	303	346	803	603	1,044
#annotations	18,210	17,767	13,090	2,154	2,487	4,874	2,453	5,343
#annotations/item	7	8	8	7	7	6	4	5
Avg. accuracy	71.4	72.7	79.2	76.3	74.1	73.1	75.8	76.0

4.4 Personalised news recommendations

Our 20 experimenters were requested to evaluate news recommendations according to 10 user profiles obtained from Flickr and del.icio.us datasets. Using News@hand and its recommendation algorithms, they had to evaluate the 5 top ranked news items for each user/topic, specifying whether a recommended item would be relevant or not for the users taking into account their profiles. Table 3 shows the average results. Each value represents the percentage of evaluated news items that were marked as relevant. The results are compared with those obtained with a classic keyword-based algorithm [4] applied to the initial folksonomy-based user profiles.

Table 3. Average relevance values for the 5 top ranked news items recommended by News@hand

	headlines	**world**	**business**	**technology**	**science**	**health**	**sports**	**entertainment**
keyword-based	46.3	34.3	39.0	43.5	35.9	21.1	58.0	33.5
News@hand	57.0	53.2	72.8	94.0	60.9	40.6	98.2	60.4

5 Conclusions and future work

The combination of folksonomy information with knowledge available in the Semantic Web is in our opinion a powerful and promising approach to provide flexible, multi-domain collaborative recommendations. It benefits from two major issues: the easy adaptation to new vocabularies, and the supervised representation of semantic knowledge. Folksonomies and Wikipedia repositories continuously and collaboratively grow, providing consensual up-to-date semantic information about user preferences and items. On the other hand, ontologies allow us to describe and organise the above information, so that relations between concepts can be defined and used by fine-grained content retrieval and recommendation strategies.

We have presented techniques that filter personal tags, and integrate them into multi-domain ontological structures considering semantic information extracted from Wikipedia. Annotating item contents with concepts of the same knowledge bases, we relate user profiles and item descriptions under a common semantic concept space, fact that is exploited by several ontology-based recommendation algorithms. We have conducted preliminary evaluations of the above techniques obtaining favourable results. However, more detailed experimentation should be done in order to obtain founded conclusions about the benefits of our proposals.

Acknowledgements

This research has been supported by the European Commission (FP6-027685 – MESH, IST-34721 – TAGora). The expressed content is the view of the authors but not necessarily the view of the MESH and TAGora projects as a whole.

References

1. Alfonseca, E. Moreno-Sandoval, A., Guirao, J. M., Ruiz-Casado, M. (2006). *The Wraetlic NLP Suite*. Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006).
2. Cantador, I., Bellogín, A., Castells, P. (2008). *A Multilayer Ontology-based Hybrid Recommendation Model*. AI Communications, special issue on Recommender Systems. In press.
3. Cantador, I., Castells, P. (2008). *Extracting Multilayered Semantic Communities of Interest from Ontology-based User Profiles: Application to Group Modelling and Hybrid Recommendations*. Computers in Human Behavior, special issue on Advances of Knowledge Management and the Semantic Web for Social Networks. Elsevier. In press.
4. Castells, P., Fernández, M., Vallet, D. (2007). *An Adaptation of the Vector-Space Model for Ontology-based Information Retrieval*. IEEE Transactions on Knowledge and Data Engineering, 19 (2), pp. 261-272.
5. Cattuto, C., Loreto, V., Pietronero, L. (2007). *Collaborative Tagging and Semiotic Dynamics*. Proceedings of the National Academy of Sciences 104(1461).
6. Chirita, P. A., Costache, S., Handschuh, S., Nejdl, W. (2007). *PTAG - Large Scale Automatic Generation of Personalized Annotation Tags for the Web*. Proceedings of the 16th international conference on World Wide Web (WWW 2007). Banff, Alberta, Canada.
7. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G. (2006). *Information Retrieval in Folksonomies: Search and Ranking*. Proceedings of the 3rd European Semantic Web Conference (ESWC 2006). Budva, Montenegro.
8. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G. (2007). *Tag Recommendations in Folksonomies*. Knowledge Discovery in Databases 2007, pp. 506-514.
9. Miller, G. A. (1995). *WordNet: A Lexical Database for English*. Communications of the Association for Computing Machinery, 38(11), pp. 39-41.
10. Niwa, S., Doi, T., Honiden, S. (2006). *Web Page Recommender System based on Folksonomy Mining for ITNG'06 Submissions*. Proceedings of the 3rd International Conference on Information Technology (ITNG 2006). Las Vegas, Nevada, USA.
11. Specia, L., Motta, E. (2007). *Integrating Folksonomies with the Semantic Web*. Proc. of the 4th European Web Semantic Conference (ESWC 2007). Innsbruck, Austria.
12. Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., Servedio, V. D. P. (2007). *Folksonomies, the Semantic Web, and Movie Recommendation*. Proc. of the 1st Int. Workshop "Bridging the Gap between Semantic Web and Web 2.0". Innsbruck, Austria.
13. Vallet, D., Castells, P., Fernández, M., Mylonas, P., Avrithis, Y. (2007). *Personalised Content Retrieval in Context Using Ontological Knowledge*. IEEE Transactions on Circuits and Systems for Video Technology, 17 (3), pp. 336-346.