Can Data Mining Techniques Ease The Semantic Tagging Burden?

Fabio Forno[1], Laura Farinetti[1], Sean Mehan[2]

[1] Politecnico di Torino, Dipartimento di Automatica ed Informatica, Torino, Italy
`fabio.forno@polito.it, laura.farinetti@polito.it`
[2] SMO, University of the Highlands and Islands, Sleat, Isle of Skye, UK
`sean@smo.uhi.ac.uk`

Abstract. The effective implementation of the Semantic Web vision is highly dependent upon the widespread availability of large collections of semantically rich resources which are trustworthy and meaningful. Since semantic classification is dependent upon complex ontologies, a recognised difficulty is the steep learning curve presented to human classifiers when attempting to utilise such ontologies. One important method to foster an increase in web accessible, semantically tagged resources is to make available tools which allow users to explore and understand relevant ontologies and to present relevant categories with which to tag new data. In this paper we investigate how an important and powerful data mining technique, Latent Semantic Indexing (LSI), might help in the design and implementation of tools that guide users in semantic tagging tasks. We applied LSI to a large portion of the Open Directory Project (ODP) catalogue, one of the largest repositories of semantically tagged resources available today. We computed statistical information concerning category relationships in the ODP data set, and we incorporated structural information by modifying the construction process of the LSI space. Using this basis, we conducted a comparative experiment where a machine generated classification of new documents was evaluated against a classification created by a group of human users. This paper includes an evaluation and discussion of the experimental results.

1   Introduction and goals

The power of the World Wide Web as a mechanism for sharing information in a globally connected network has impacted on wide sectors of society, but methods to increase the effectiveness of the Web are required which should combine machine understandable content and machine reasoning capabilities with the data. To meet the need, researchers have called for the next generation of the web, the Semantic Web (SW) (Berners Lee at al., 2001) , where web information objects transform into web knowledge objects.
In this new web generation, web documents will not only be designed for human reading, but also for machine processing. Additional knowledge will be available to provide context and relationships about information objects. The SW will be

built on three components: structured collections of machine understandable information, inference rules with which to conduct machine reasoning, and software agents able to process information and exchange results with other programs.

The most commonly envisioned SW implementation of these components requires: a systematic, shareable, computer-oriented representation of the world (often referred to as an ontology), semantic annotations of web resources, and software agents to retrieve and manage knowledge instead of unstructured data. Many languages for semantic tagging of resources can be found and this has been an active research area (Gomez-Perez and Corcho, 2002), however, to date, very few large-scale *(semi)-structured* collections of resources have used these languages. Examples of such a large scale collection are the Open Directory Project (ODP, 2003) and news and weblog syndications through RSS (Winer, 2002). We feel that this lack of large scale implementation is one of the reasons why available search engines are far from being "semantic", but rather still use a "keyword" approach. ODP, in particular, is based on the efforts of a small community of users which has produced and currently maintains a general ontological structure in which web resources are manually annotated in RDF. The manual annotation of resources, as a constraint, results in only a few documents, 3 million at present, being semantically tagged compared to the the total web document population; ODP currently holds some 3 million tagged resources out of a conservative estimate of 4 billion documents on the web, meaning the ODP collection is less than 0.075% of the total web population.

A trivial solution would be to enlarge the community of annotators, leading to decentralised management, but this would lead to problems related to trust and to the usage and maintenance of ontologies. The first problem concerns the trust of annotations actually matching between document content and document semantic tags; failure in this regard could lead to spamming, among other problems. The proposed SW architecture addresses this problem, and proposes the creation of a "network of trust", but Tim Berners-Lee himself (2001) foresees the stage of "trusted web resource" only being established after year 2010. The second problem is a practical issue, concerning a common agreement and understanding of ontologies in use by a large community of annotators, which often have very specific resources to deal with. Automatic or even semi-automatic tagging of resources can be proposed to counteract these issues, yet large-scale semi-automatic tagging of resources is still far from being a reality, mainly due to a lack of user-friendly tagging tools that are effectively linked to ontologies.

In this paper we start with a description of the some of the observed problems that human classifiers have when trying to catalogue a web resource, especially when they are expert in the site domain but not generally expert in classification tasks and are not familiar with the domain ontologies, which often are broader than their direct knowledge of the subjects. This scenario fits well with a highly likely future scenario of the SW, where people will catalogue their own web resources using ontologies designed by others.

For this purpose, after a short introduction to the context of our research in section 2 and a review of related works found in the literature in section 3, we

describe an experiment we conducted to empirically measure the difficulties involved in manual semantic tagging, where the tagging is oriented to document classification for information retrieval and not simply for adding annotations useful to human readers planning to share information. This section includes discussion of the experimental results and argues for the need for an automatic tool capable of easing the semantic tagging burden. Such a tool would allow for convenient browsing of the ontological elements and propose a list of suitable categories from among which to select.

The paper then considers the potential of applying data mining techniques to semantic tagging tasks, and in section 4 we propose preliminary results of an approach that applies a powerful and well assessed data mining technique, Latent Semantic Indexing (Deerwester at al, 1990), to the ODP catalogue. This is done in order to extract information concerning the match between document content and selected ontology categories. The proposed tool can exploit this information for extending the cataloguing to a much larger number of documents, through a learn-by-example approach, proposing the best match ontology elements to human classifiers as part a semi-automatic tagging process. Finally, in this section, a preliminary test of the designed tool is described and discussed.

We conclude the paper with a brief section summarizing our conclusions in section 5.

2   Context of research

The literature has numerous reports of efforts and experiments which extract information from unstructured data; this extracted information is then used to subsequently build some semantically meaningful elements with which to tag the indexed resources. These efforts in fact span over more than 20 years of efforts to extract classification information from such unstructured data sets. Many modern methods use various machine learning techniques, including feature vector representations similar to the approach used in our experiment. For example, in one series of experiments (see Grobelnik et al., 1998), a naive Bayesian classifier was used on text data derived from three domains in the Yahoo hierarchy; from this an n-gram feature-vector document representation was constructed and application of this lead to a high correlation between the human built classification and the classification predicted by the machine learning algorithm. Another approach (Li et al., 2001) was to use a machine learning technique which analysed natural language sentences in documents and annotated, using RDF tags, each prime sentence in the document. This was based on semantic analysis on these natural language sentences using Conceptual Graphs. The major distinction between these kinds of approaches and our own is that these approaches are targeted at automatic classification of existing and new resources whereas our overall aim is to provide a supporting tool to allow human users to use ontologies to semi-automatically tag existing and new resources. Vargas-Vera et al. (see Vargas-Vera et al., 2001) describe a semantic annotation tool for extraction of knowledge structures from web pages through the use of simple

user-defined knowledge extraction patterns. The semantic annotation tool they built contains an ontology based mark up component which allows the user to browse and to mark up relevant pieces of information. Also it contains a learning component which learns rules from examples and an information extraction component which extracts the objects and relations between these objects. Another example is the CREAM framework (see Handschuh et al.,2001). CREAM (Creating RElational, Annotation-based Metadata) is a framework for an annotation environment that allows construction of relational metadata, i.e. metadata that comprises class instances and relationship instances. These instances were not based on a fixed structure, but on a domain ontology. This approach, similar to other similar annotation efforts which are being pursued, differ from the first class of applications in that they are human centric. They do differ from our own approach in that a major aim of these systems is to populate extendible ontologies through the use of the information extraction component, whereas our own is to use an ontology to guide user selection of semantic tags.

We decided to apply LSI to the ODP, since it is a well assessed technique widely described in information retrieval literature. Since its potentials and limitations are well known, it is, therefore, easy to understand the added value of its application to already semantically structured data as opposed to flat corpora. Among the several web directories available on the net, e.g. Yahoo!, we selected the ODP catalog, since it is the only resource annotated with RDF and thus suitable for automatic processing. ODP data is available for download in two large RDF files containing respectively structure definitions and URL annotations. Like the other web directories, ODP can be labelled as a taxonomy with more accuracy than as an ontology, since it does not define strict semantic relationships between nodes, but rather describes generic topic connections (e.g. "Subtopic", "Related") as a graph. URLs are then annotated with one or more topics and a textual description. Though this configuration does not allow agent based reasoning yet, due to its weak semantic components, we believe it provides a good example of the benefits that could derive from semantically structured data in general.

3   Manual tagging

3.1   Experiment description

The goal of the experiment was to replicate and measure the difficulties that humans have when trying to semantically classify web resources. We believe that the real exploitation of the potential of the SW will only happen when large numbers of web resources are semantically tagged, and that this requires that many humans be involved in the tagging task. This is true because many more people are creating web resources who do not have expertise and experience with classification and cataloguing documents than those who do have such skills. Therefore, in our experiment, we purposefully involved people that have a good general knowledge of a subject domain but have no previous experience in cataloguing resources or using ontologies. This profile, in our opinion, fits one

of the largest user groups which will require support in the new SW.

Our subjects were graduate students in the fields of engineering, economics and social sciences; 21 students participated in the experiment, and had to perform four tasks as well as completing a questionnaire after each task. They each had to complete and return the questionnaire after each task before going to the following one, to prevent their results being used in the following task.

Task 1 The subjects had a list of web sites to visit, the topics of which were wide ranging; They included sites for computer science, social science, astronomy, science fiction, ethnographic studies and humor. The students had to record the three keywords that, in their opinion, were the most appropriate match to the subject of each site. No list of keywords was provided, so they were completely free to choose any keyword. They also had to provide a written explanation about the strategy they used in performing this task.

Task 2 For each of the 11 web sites previously considered, the subjects had to find the best matching category among the ODP categories. In this task, however, they could only browse the categories and sub-categories without looking at the categories' descriptions nor accessing already catalogued documents as examples. Therefore they could only use category name and the path from the root to the category to determine semantic information about a category. Again, they also had to provide a written explanation about the strategy they used in performing this task.

Task 3 This task was similar to task 2; for each of the 11 web sites the subjects had to find the best matching category among the ODP categories. This time, however, they had access to all the ODP information about the category meaning and could look at the documents catalogued in the categories as examples. Yet again, they had to write a few words about the strategy they used in performing this task.

Task 4 In this task, the students had to record the main difficulties they had experienced while cataloguing the web resources in general, and in performing the previous three tasks in particular. We used this information together with the other collected data to interpret the experimental results.

The following section reports the experimental results and attempts to interpret them, providing an understanding of the difficulties that humans have in cataloguing web resources when they are not classification experts, when they are not familiar with ontologies and when they have no clear guidelines for performing these kinds of tasks.

3.2   Results interpretation

Since we want to provide both a quantitative and a qualitative interpretation of the experimental results, before proceeding with the analysis we try define some numerical quantities in order to provide an objective data framework. However, in general, numerical quantities relating to the process of manual tagging are not used, therefore we first introduce the indices we have used, summarized in table 1. Each table row refers to a web site (11 in total) which has been classified by the subjects that took part in the experiment, while the last row contains the averaged data.

Total keyword number The first three columns represent the total number of keywords used by the subjects in tasks 1, 2 and 3 respectively; they represent a first measure of the dispersion introduced by manual tagging. In tasks 2 and 3, users had to specify one ODP category using the complete path from the root of the ontology, i.e. */Science/Technology/Education/*. In order to be able to compare the dispersion of keywords with task 1, we split the full path into its components, thus obtaining a set of keywords from each topic, e.g. (*Science*, *Technology* and *Education* for the previous example).

Keyword intersection Columns labeled *I12*, *I13*, *I23* contain the number of keywords in common between the sets collected in the different tasks. For example, row 1 states that the keyword set resulting from task 1 has only 7 keywords in common with task 2 and only 6 keywords in common with task 3 respectively, while the latter tasks have an intersection of 19 keywords.

Entropy The last three columns, *H1*, *H2*, *H3*, represent the entropy of each set of keywords in the three tasks. In order to calculate the entropy of a set of keywords we have considered each keyword as a code symbol of a language, and calculated its frequency $p_i = n_i/N$, where n_i is the number of occurrences of the keyword and N the total number of occurrences. Then we have applied Shannon's definition of entropy of a codebook (1963):

$$H = -\sum_i p_i log_2(p_i)$$

Information Theory interprets H as the total amount of information a codebook can carry. Physics interprets H, as in statistical mechanics, as the measure of the disorder of a system; importantly, the two interpretations are not in conflict (Ott, 1993). If we consider a set of symbols as growing more ordered as any particular symbol's frequency increases, then we may also interpret this as less information overall contained in the set from an information theoretic standpoint, with both of these perspectives being labelled as low entropy in the set. Hence we can use H as a measure of the degree of agreement between tags selected: high levels of agreement between selected keywords produces a smaller set of unique keywords, or a codebook with low H; Alternatively, low agreement between keywords selected results in a large set of unique keywords, or a codebook with high H. We choose to

use H, rather than n_i, because it is a more meaningful measure. A codebook with a large N, but a small number of them recurring with high frequency, is more ordered than the same codebook with symbols recurring with equal frequency.

Keyword distribution Continuing with the codebook analogy, in figure 1 we plot the keyword recurrence frequency distribution obtained in each of the three tasks of the experiment. For each web site we have sorted the set of keywords in reverse order according to their frequency. The distribution magnitude relates to the size of the core of a keyword set, i.e. the set of the most commonly recurring keywords according to a parameterized metric.

Table 1. Keyword total number, intersection and entropy for each task of the experiment.

Site	T1	T2	T3	I12	I13	I23	H1	H2	H3
1	34	22	22	7	6	19	4.39	3.76	3.81
2	36	28	33	9	10	20	4.55	4.22	4.52
3	28	27	30	7	10	18	3.98	4.02	4.31
4	34	17	7	5	4	7	4.44	3.08	2.53
5	38	20	19	4	4	14	4.50	3.43	3.43
6	40	15	14	4	8	7	4.66	2.83	2.58
7	26	26	22	10	10	16	3.70	3.80	3.44
8	40	37	32	10	11	24	4.58	4.80	4.57
9	28	29	23	7	5	20	3.75	4.17	3.60
10	38	24	24	3	4	17	4.45	3.96	3.88
11	28	44	38	6	6	33	4.21	4.65	4.38
Avg	33.63	26.27	24.00	6.54	7.09	17.72	4.29	3.88	3.73

The first characteristic of the tagging process determined from the indices is the high number of keywords used by the classifiers in the three tasks. In the first task, where students freely selected three keywords describing the web sites, there was an average of 33 unique keywords used to describe a site. This means that the each student chose, on average, 1.5 unique keywords to describe the same site. A high keyword dispersion level was foreseeable and our data supports the hypothesis that classification is a highly subjective task. More interesting was the relatively high dispersion of keywords in tasks 2 and 3. The total number of selected keywords decreased, but each user still contributed more than one new keyword to the set. This fact could be interpreted in the following ways:

- the use of ontologies is of little help for users who have little experience in classification tasks;
- in certain situations ontologies may represent an obstacle for users (see, for example, row 11 in table 1) in that when they start looking for a suitable category, they have already mentally selected a classification, perhaps even
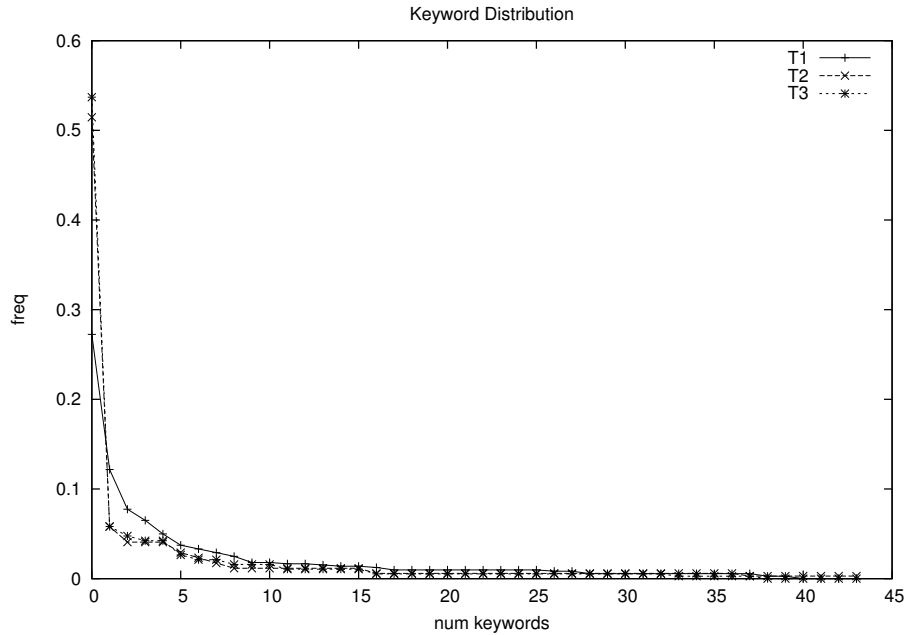
Fig. 1. Keyword distribution according to usage

    subconsciously. Subsequentially, they are influenced by their own predetermined category and are unable to find a matching category;
- the ODP ontology is a wide and dispersive taxonomy, with the aim of classifying almost everything present on the web: this may lead to incongruencies, and also to classification areas with inappropriate granularity.

    These results are not surprising, since ontologies have been designed by experts to facilitate automated reasoning, not to ease the human tagging burden. We may even theorize that the more an ontology is specialized and finely grained, the more human taggers will have difficulties in selecting the correct categories from it. Therefore, there is an evident mismatch between the need of precision and specialization of ontologies destined for machine reasoning, and the ability of humans to utilize them.

Another indication of the mismatch between the human classifier's mental processes and the ontology they are forced to use may be derived from our intersection indices. In average only 6 keywords out of the 33 selected in task 1 have been picked also in task 2, less than 20%. This percentage is similar but slightly higher to the intersection between tasks 1 and 3. If confirmed with a larger number of test cases, this increase could be interpreted that when given more time to explore the ontologies, human classifiers are inclined to look for categories matching their first choice.

Comparing tasks 2 and 3 we note two more trends: the small decrease of keyword

dispersion and the higher intersection of chosen keywords. Possible interpretations include:

– the diminution of the dispersion may suggest that users, with adequate support to navigate the ontology, may choose the appropriate categories more easily, with less frustration during the process;
– examples of classification, which are the only additional support that users have when moving from task 2 to task 3, are not sufficient to improve significantly the similarity of classification outcomes; as confirmed by user comments, support tools for ontology exploration could lead to more agreement in tagging; lack of concordance, in fact, seems to be not due to low user concept comprehension, but rather to user difficulty in acquiring a broad understanding of the ontology.
– during tasks 2 and 3 users may have explored only a small part of the ontology, and in both cases they learn to not explore it further; this is another indication that there is a need for tools assisting in the navigation of ontologies.

The analysis of the H_i supports the keyword scattering we have already observed. The calculated entropy in all the tasks always remains high. If all the human classifiers had selected the same three keywords, the H would equal 1.58; with a completely random keyword selection, i.e. all keywords are unique, H would equal 5.97. The average values in table 1, 4.29, 3.88 and 3.73, testify to the high degree of dispersion in all the three tasks (since H is logarithmic, these values are significantly different).

With respect to figure 1, we find that, even if the keyword dispersion is high there is a very small set of keywords (one or two) that nearly half of the subjects identify in all the tasks. This is reasonable for tasks 2 and 3, since the sets must always contain the same root keywords. It is more interesting to remark that also in task 1 there is a kernel of common keywords; as a consequence we may imagine that tools suggesting categories of the ontology starting from keywords expressed in free form by users may facilitate the tagging process and lead to more concordant results.

The analysis of the last experimental task supports some hypotheses in our model describing the difficulties involved in manual tagging, and offers a qualitative description of the problems from the point of view of human classifiers. Here we summarize the main difficulties the subjects reported encountering:

– the deep structure and the presence of many internal links connecting different tree branches allows the user to become easily lost in the structure;
– many categories are too semantically similar and classifiers have difficulties in selecting just one topic;
– category names and associated keywords sometimes do not provide enough hints about the contents, thus hindering navigation;
– some classifiers complain that the ontology is not exhaustive, but we think rather that the correct interpretation is given by other classifiers reporting that the ontology does not cope with their own personal classification.

Several classifiers also described the strategy that they adopted in the process:

- when exploring the ontology users start from the task 1 keywords;
- the exploration usually proceeds depth first;
- the presence of the *related* topics links also allows some exploration in breadth;
- some users autonomously followed an approach similar to our proposed method, using the Google search engine with the keywords found in task 1 in order quickly find some matching categories; therefore, we think that the support of tagging tools operating on user suggested keywords will be intuitive to use.

In summary, classifiers often start from a set of keywords they choose independently from the given ontology. This set derives from their own experience with the web site topic and from their personal usage of the language, and thus is very likely to *not* match with the terms present in the given ontology. Then they explore the ontology in order to find the best match for their own keywords, but they encounter many difficulties due to the ontology topics not matching well with their own, preselected topics and also due to difficulties encountered while exploring the ontology.

These last conclusions support our hypothesis that a tool is needed which is able not only to facilitate users browsing ontologies (which, in itself, would provide a good level of support), but also able to automatically suggest a restricted list of categories from which the human classifier can select the best one.

4 Automated support for manual tagging

4.1 Latent Semantic Indexing

We first give a brief introduction to LSI. In information retrieval a query is usually expressed in a vector space (Salton, 1975), where vectors represent documents and the vectorial components formed from the document constituent terms. Given a similarity measure, the document vectors best matching a query vector can be selected. LSI operates in a similar fashion, except that a transformation to the $term \times document$ space is achieved through the application of a Singular Value Decomposition (SVD). A corpus of n documents with m indexing terms is represented by the matrix $A^{(m \times n)}$, where each element a_{ij} represents the weight of ith term in document j. Several weighting schemes for term weighting have been proposed in the literature1996 and LSI normally uses a functional combination of local and global weights; thus $a_{ij} = L(i,j) * G(i)$, where $L(i,j)$ is the weight of the ith term in document j, usually its frequency, and $G(i)$ is a function of its global weight in the corpus, usually term entropy dependent upon inverse document frequency. Subsequently the truncated SVD is applied to the matrix A, yielding:

$$A_k = U_k \times \Sigma_k \times V_k^T$$

where $\Sigma_k \equiv diag(\sigma_1 \cdots \sigma_k)$ is a diagonal matrix, with the k largest singular values of A, and $U_k \equiv (u_1 \cdots u_k)$ and $V_k \equiv (v_1 \cdots v_k)$ are the associated left and right singular vectors respectively. In (Berry, 1995 it is shown that A_k is the best approximation in k dimensions of the original matrix A. The result of this transformation is that the original vector space model is folded into a much smaller subspace, which should better summarize word and document relationships, capturing hidden *semantic* links in smaller feature vectors (Berry et al., 1995, Hoffmann, 1999, and Papadimitriou et al., 1998).

In the LSI space, queries are expressed as in the original vector space, where, given a vector q representing the set of query terms, each document score is given by $s_i = q^T x_i$ (x_i is the feature vector of document i. In LSI the query vector is simply projected into the k subspace, via: $s = (q^T U_k)(\Sigma_k V_k^T)$.

4.2 Application of Latent Semantic Indexing to the Open Directory

Our tagging tool is based on the application of LSI to a large portion of previously tagged resources in the ODP; specifically, we selected all the Science, Computers, Arts, Society and Recreation subtrees, resulting in a collection containing more than 50,000 topics and 1,000,000 URLs. In this ontology we find two distinct types of resource: (i) topics organized in a tree-like structure, all having a short textual description of the content, and (ii) links tagged with one or more topics and a brief description. LSI is normally applied to flat document corpora, without any structure; since the associated vector space model does not take into account existing semantic relations and not lose this critical information, we had to make some changes to the process whereby the $term \times document$ matrix is constructed. In our experiment we have built several vector spaces following different methods, in order to evaluate the best ways to incorporate structural information.

In all matrices we have used the same scoring function for terms:

$$a_{ij} = f_{ij} \cdot G(log_2(idf(i)), logMean)$$

where f_{ij} is the ith term frequency in the document j, and the second term gives its global score, computed as i^{th} term inverse document frequency (idf) and weighted by the Gamma distribution centered in $logMean$. $logMean$ is a parameter dependent upon the size of the document collection. In this modified weighting scheme, based on Salton (1975), our previous experiments outperformed entropy based measures, since the modified scheme penalizes terms recurring with a very high frequency (stopwords were already pruned in document preprocessing).

In the following we describe the four different methods we used to attempt to incorporate the structural information into the LSI spaces, and we shall refer to these as $A_k^{(1)}$, $A_k^{(2)}$, $A_k^{(3)}$ and $A_k^{(4)}$, respectively. In $A_k^{(1)}$ we considered topic descriptions as documents, enriched with the descriptions of contained URLs. Since the quantity of associated text was not large, the matrix scaled well; in fact, with this strategy it is conceivable to process the entire ODP with consumer level hardware.

In $A_k^{(2)}$ we inserted as documents both topic descriptions and text extracted from indexed URLs. In this case, in order to scale to the dimension of our experiment, we have applied a random selection of documents and terms, as normally performed in LSI.

In $A_k^{(3)}$ we used only topic descriptions again, similar to $A_k^{(1)}$, but we started incorporating structural information. We made the assumption that, since the ODP has a tree-like structure, a topic is qualified not just by its description, but also by the description of its child nodes. Thus, in building $A_k^{(3)}$, we associated to each topic also the text of its children (we ignored *related* topics and *symbolic* links). The resulting matrix was less sparse, but since for our SVD implementation the occupied memory is given by $k \cdot min(m, n) + nnzeros$, where k is the LSI dimension and $nnzero$ the number of non zero values in the matrix, and also, since the first addend usually grows faster, this approach continued scaling well.

In our last matrix, $A_k^{(4)}$, we used text from indexed documents, but we employed a different method from random selection. Random selection, in fact, is based on the assumption that we do not know if there is a structure in the indexed corpus, thus a completely random choice of the documents and keywords is the best guess we can make in order to obtain a representative subset. On the other hand, ODP catalogued URLs are already grouped by topics, and they form a bunch of small clusters of manually selected examples for each category. We exploited this information, extracting the core of most frequent words for each topic, and we then used this new lexicon as the associated text for the topics. We weighted words with their frequency in each set, and we recomputed their global weight according to the new lexicon. Using the codebook analogy described in section 3, and performing some tests, we have established a heuristic that keeping the words contributing to the central 75% of the associated cumulative distribution function is a good compromise between keyword set size and retained information. Without a given structure, this selection would have been ineffective or even counterproductive, since we would have retained only common usage words and not topic representative keywords.

4.3   A preliminary experiment and discussion

In the experimental phase we have evaluated two distinct aspects of our LSI semantic tagging tool, namely: (i) whether the incorporated structure information improves LSI performance in terms of recall and precision (Raghavan, 1989, and (ii) whether it can be exploited as a support for manual tagging.

Benchmarks on text retrieval performance, in terms of precision and recall, are usually performed on a fixed set of queries on standard corpora. To assess the capability of our LSI implementation in capturing deal with structural information, we defined a set of queries composed of few keywords and spanning several ODP terms and we applied them to all the $A_k^{(i)}$. During the evaluation of the results we had to consider that the ODP structure presents many overlaps, and the same topic may be present in different branches (e.g. computational linguistic resources are present both under */Computers/Artificial_Intelligence/* and

/Science/Social_Sciences/Languages_and_Linguistics/). Therefore, in contrast to a simple count of returned pertaining topics (leaves) which would not have resulted in significant understanding, we concentrated mostly on the relevance of the returned major topics (internal nodes).

Evaluating query results, we found that both $A_k^{(3)}$ and $A_k^{(4)}$, built using structural information, outperform the corresponding $A_k^{(1)}$ and $A_k^{(2)}$, built with flat corpora in both precision and recall. Specifically, $A_k^{(1)}$ and $A_k^{(2)}$ behaved well for certain topics and completely missed others. The reason for their irregular behavior must be investigated further, but we believe it may derive from the lack of sufficient representative keywords for some topics; with random selection, the method results in topics with more catalogued resources being better represented, since their elements have more chances to be selected. In $A_k^{(3)}$ and $A_k^{(4)}$, we improve the selection of meaningful keywords, since they are either inherited from child nodes ($A_k^{(3)}$) or selected using topic aggregation of resources ($A_k^{(4)}$). Finally matrix 4, besides scaling very well in terms of memory occupation, since we can considerably reduce the lexicon size, scores a better relevance then matrix 3. We think that the worse performance of matrix 3 might be found in the heterogeneity of document lengths. Its building process, the inherintance of describing text from child nodes, in fact, produces short documents for nodes deep in the structure, and very long ones for nodes near to the root. Since the behavior of LSI is not well understood when document length has a variance of magnitude orders, we believe that the vector space represented by matrix 3 needs more investigation.

We then evaluated whether obtained results could be used in order to help manual tagging. The main issues which emerged in the 3.2 related to user difficulties with their orientation in the ODP structure, as opposed to the need of exploring it in depth or breadth. Thus, human taggers could exploit LSI derived information, in order to obtain a few meaningful suggestions as starting points in the ODP structure. There are two possible approaches: (i) using documents to be tagged as queries in the LSI space, (ii) allowing user to express some keywords, or a short document description and use these as queries. Returned results would then be sent to users as possible starting points for ODP exploration.

We experimented with both methods on our $A_k^{(i)}$, obtaining promising results. As an example, in table 2 we list the suggested categories using strategy (i) for one of the documents we used in the experiment of paragraph 3. The document is the homepage of L. M. Krauss, the author of books about science in science fiction, where it can seen that the suggested categories are very relevant. In contrast, the same document gave several problems to manual taggers, since they looked for Star Trek (the main topic of the books) under the 'TV Shows' area, and they had difficulties in finding relevant topics (alternative science) under the 'Science' section.

Our experiments demonstrate that queries composed of a few keywords usually matches to relevant categories more efficiently than using only documents themselves. This may be due to the fact that the few keywords should give a better

definition of a document in the LSI space rather than the entire document, since many contained words may be considered *noise*. We observed also that intermediate nodes of proposed topics are the best candidates to suggest to users as starting points. Deep topic nodes, in fact, can suffer from some noise that LSI has not been able to eliminate and their precision may be low; in contrast the precision of intermediate nodes in the results in performed queries is very high.

Table 2. Some of the suggested categories for the homepage of L.M. Krauss.

/Science/Anomalies_and_Alternative_Science/Astronomy,_Alternative/Cosmology/
/Science/Social_Sciences/Ethnic_Studies/
/Science/Physics/Relativity/People/Hawking,_Stephen/
/Science/Math/Applications/Mathematical_Economics_and_Financial_Mathematics/
/Science/Astronomy/Education/
/Science/Astronomy/History/People/Kepler,_Johannes/
/Science/Physics/Alternative/Superluminal_Physics/
/Science/Earth_Sciences/Geology/Geochronology/Radiometric_Dating/
/Society/Future/Predictions/Scientific/
/Society/History/By_Topic/Science/

5   Conclusions

For the SW to become a reality, there needs to be an extensive and rich set of web objects which have associated semantic information. It is generally accepted that there will be common repositories of domain specific semantic information, namely those collections which are generally referred to as ontologies within the relevant communities. This paper has identified and analysed some of the problems that educated but non-expert users experience in trying to apply existing ontology information when classifying and tagging data resources. Using the ODP, we observed a number of users completing classification tasks with ODP categories.

The results of our user experimentation indicate that, even with access to structured ontological knowledge, the classifiers chose many different resource descriptive tags from one another, resulting in a high level of dispersion amongst the final tag set. It seems that ontologies are not immediately useful to experienced users who do not have extensive experience with classification tasks. Most normal users have a preconceived term in mind when tagging a resource and when this preconception does not map to the ontology, conflict often occurs.

It is reasonable to speculate that the more detailed the ontology, the more serious this conflict with normal users of the ontology can be. Complex ontologies seem to lose users who are trying to navigate them, and this leads to demotivation on the part of the manual classifier to explore the ontology fully. Users

often complain of lack of completeness in ontologies, and we interpret this as a clash between the ontology conceptualization and the user's predetermined conceptualization in the classification task. Classifiers seem to start with a set of terms from their own experience and language registers, independent of those found within the ontology.

These difficulties support our hypothesis that a tool is needed which is able to ease the manual semantic tagging burden by the widest population of content creators when using standardized ontological resources. This tool should help users to explore ontologies, but also should automatically suggest a restricted set of tagging terms drawn from the ontology to help guide the user in their classification tasks. We have experimented with a number of different applications of LSI to ODP tagged resources. Importantly, we have experimented with new methods of incorporating semantic structural information from the initial vector space into the transformed LSI space and we have tested whether the incorporated information improves querying performance and whether in can be exploited as a support mechanism for manual semantic tagging. We found that through the incorporation of this structural information, we were able to improve both query precision and recall. This results in more meaningful selection of keywords from the ontology, which would result in more effective semantically tagged resources. In pursuit of this, we experimented with two different approaches, namely to use documents which were to be tagged as query vectors in the $A_k^{(i)}$ space, or allowing classifiers to supply some of their own preconceived terms as keywords or a short description and constructing query vectors from these. LSI seems to bridge the gap between the ontology terms and terms preconceived by human classifiers. LSI projects in the same space $term \times documents$ relations which were learned already from tagged documents using the ontology and those terms in the mind of the original ontology compilers.

References

Berners-Lee, T., Hendler, J.: Scientific Publishing on the "Semantic Web". Nature Webdebates, http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm, April 2001.

Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 5/01, May 2001.

Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using Linear Algebra for Intelligent Information Retrieval. SIAM Review 37(4):573–595, 1995.

Deerwester. S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407, 1990.

Gomez-Perez, A., Corcho, O.: Ontology languages for the Semantic Web. IEEE Intelligent Systems, Volume 17(1):54-60, Jan/Feb 2002.

Grobelnik, M., Mladenic, D.: Efficient text categorization. Text Mining workshop on the 10th European Conference on Machine Learning ECML98.

Handschuh, S., Staab S., and Maedche, S.: CREAM- Creating relational metadata with a component-based, ontology-driven annotation framework. Proc. of ACM K-

CAP 2001 - First International Conference on Knowledge Capture, Victoria, BC, Canada, October 21-23, 2001.

Hoffman, T.: Probabilistic Latent Semantic Indexing. Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval (1999) pp. 50-57.

Li, J., Zhang, L., and Yu, Y.: Learning to Generate Semantic Annotation for Domain Specific Sentences. K-CAP 2001 workshop on Knowledge markup and semantic annotation, Victoria, BC, Canada, 21 October, 2001.

Ott, E. Entropies. 4.5 in Chaos in Dynamical Systems. New York: Cambridge University Press, pp. 138-144, 1993.

Open Directory Project. http://www.dmoz.org. 2003.

Papadimitriou, C.H., Raghavan, P., Tamaki, H., and Vempala, S.: Latent Semantic Indexing: A Probabilistic Analysis. In Proceedings of the ACM Conference on Principles of Database Systems (PODS), Seattle, 1998.

Raghavan, V. V., Jung, G. S, and Bollmann, P.: A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. ACM Transactions on Office Information Systems, pages 205–229, July 1989.

Salton G., Wong, A.and Yang, C. S: A Vector Space Model for Automatic Indexing. CACM, Vol. 18, No. 11, 1975, 613-620.

Salton, G. and Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Technical Report TR87-881, Department of Computer Science, Cornell University, 1987. Information Processing and Management Vol.32 (4), p. 431-443, 1996.

Shannon, C. E. and Weaver, W.: Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1963.

Vargas-Vera, M., Motta E., Domingue S., Buckingham Shum S., and Lanzoni M.: Knowledge extraction by using an ontology-based annotation tool. K-CAP 2001 workshop on Knowledge markup and semantic annotation, Victoria, BC, Canada, 21 October, 2001.

Winer, D.: RSS 2.0. http://backend.userland.com/rss, Aug 2002.