Topical Structure Discovery in Folksonomies

Ilija Subasic and Bettina Berendt

Katholieke Universiteit Leuven, Dept. of Computer Science,
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{Ilija.Subasic, Bettina.Berendt}@cs.kuleuven.be

Abstract. In recent years social bookmarking systems (tagging systems) became one of the highly popular applications on the Internet. The main idea of social bookmarking is to organize content in a loose fashion by allowing users to completely freely annotate content. This work presents a way of combining the information retrieval (IR), Semantic Web and social web approaches of searching the Web by including general topic categories as a part of tagging systems. In this way semantic and social web are presented in a unified framework of search and indexing content. The work also shows a way of ontology learning by creating a hierarchical network of tag associations. This network is created using association rules discovery. In order to enhance these networks, IR search engine results are used to evaluate relevance of resources to a given topic. Networks of association, created by application of a modified Apriori algorithm, are evaluated with topic networks from the Open Directory Project (www.dmoz.org).

1 Introduction

Much of the recent growth of digital content is an effect of an increased number of internet users and their interest in on-line publishing. However, search for appropriate content remains one of the main problems. Early approaches to search that include information retrieval based search engines and web directories have almost completely been replaced by a second generation of search agents that beside the content itself take into account the interaction of users through, for example, PageRank or tagging. Social bookmarking tools/systems (SBSs) are one of the ways this second generation of agents tackle the search problem. SBSs enable users to express their reflection upon some content that is available on the web. The main idea of the social web through the use of social bookmarking is that a form of content organization emerges when users are allowed to completely freely annotate content. Databases of such systems consist of millions of users, tags and resources. A combination of these 3 objects (users, tags and resources) constitutes a folksonomy [20]. SBSs rely on the principle of collective intelligence. This principle, older than the web itself, is based on the idea that every decision made by a group of people with diverse expertise is better than the decision made by a single domain expert. Based on a different organization of SBSs they could be divided into self-tagging, which allow only content creator to tag, and non-self-tagging systems, which enable any user to tag any content [8]. In this paper's case study, we consider a non-self-tagging system. Nevertheless, the approach could be applied to self-tagging SBSs as well.

The problem faced by most SBSs is how to utilize the huge metadata repository created by users. Most applications just use tag clouds as a representation of their metadata. A tag cloud is a statistical overview of tagging behaviour of users. This way of summarizing folksonomy only offers users simple search based on the frequency of keywords (tags), without taking into consideration content or context of a resource. The analysis of tag distributions has shown that they follow some structure [16]. Nevertheless, most of the applications that enable users to tag provide little or no structured way of presenting tags to the users.

The need for structure discovery in such large repositories must involve some level of intelligent data analysis. This mix of user-generated (meta)data and intelligent ways of organizing and presenting it to users could be regarded as the next generation of web search. The main idea of this approach is that users' contributions are combined with some 'ground truth'. As ground truth, we consider content organization by predefined categories done by some agent (human expert or machine learning algorithm).

The key motivation of the present paper is to show a possible way of using data mining methods to benefit both from social and Semantic Web. The paper shows the use of data mining for discovering structure inside folksonomies by applying an algorithm for topic specific association rules discovery. Its goals are answers to three main questions:

1. Are resources tagged with some tag similar to those that one can find using IR-based search engine using the tag as query?
2. Can networks of tags and associations between them approximate existing topic hierarchies?
3. Does topic relevance inclusion in analysis enhance this approximation?

The first goal is thus to find underlying topics in folksonomies, and the second is to compare the topic structure. Building on this, the third goal investigates the inclusion of resource-to-topic relevance into structure discovery. The idea of topic relevance is that user input has to satisfy some conditions to be included into the structure discovery process. An algorithm for topical structure discovery in folksonomies is developed to meet this goal.

The paper continues by giving a related work overview in section 2 and then describes the topical structure discovery algorithm in section 3. Section 4 describes a case study and evaluation results, and the last section gives a brief conclusion of the presented work.

2 Related work

To be able to apply data mining methods to folksonomies, one must define them formally. We adopt a model described in [13], which models a folksonomy F by its representing tripartite hypergraph:

$$H(F) = < V, E > \quad (1)$$

where *V* is a set of nodes that is the union of the sets *A*,*C*,*I*, which stand for users (identities), tags (concepts) and resources (items), and *E* is a set of edges denoting which user tagged a certain resource with a certain tag: $E = \{\{a,i,c\} \mid (a,i,c) \in F\}$.

2.1 Structure discovery in folksonomies

In [9], the authors examine the way tag clouds are created and evaluate resource overlap using three different measures which express the weight of the tag inside a tag cloud. The authors compared different weighting schemes using the Jaccard index to calculate the relative co-occurrence of tags in two different resource sets. Although very popular and used by most social bookmarking systems, tag clouds are just ways of summarizing basic statistical properties of a tag set. Therefore, different data mining methods have been applied in order to describe folksonomy structure in a better way.

From the graph described in equation (1), the author [13] derives different graphs. For tag clustering, he extracts an IC (item-concept) graph, a bipartite graph of resources and tags, on which a graph clustering algorithm is applied. Alternative clustering approaches [4, 20] are modularity and EM clustering. The application of EM clustering showed that a small number of concepts (40 for del.icio.us) can be used to group tags. However, clustering just shows that there is a relation between tags (in the similarity sense), but not the nature of this relationship.

Another way of discovering links between tags is the application of association rules discovery algorithms [14]. The goal is to learn association rules from a folksonomy. The quality of a rule is measured by support and confidence values of tags describing an item. Although folksonomies should in a way present an alternative to creating taxonomies, the hierarchical nature of concept relations should not be discarded in explorations of hidden structures inside folksonomies. A way of creating taxonomic structure from a set of association rules is described in [15]. It consists of the creation of a graph whose vertices are a set of tags that form the association rules, whose edges are connections between tags which appear in the same rule (directed like the association), with edge weights given by the confidence of a rule. To get a hierarchical organization, only the edges with maximum weight are kept. Hierarchical relationships between tags were also mined from tagged resources based on the similarity of resources in tag vector space in [10].

The disadvantage of all of these methods is that they do not incorporate the content and context of tagging. By content, we mean the content of the tagged resources; by context, we denote the combination of tags applied by a given user to an item.

2.2 Tagging behavior

To understand which kind of structure can be mined from SBSs, the behaviour of users in such systems must be understood. An overview of SBSs and the incentives for their use is given in [8]. The investigation of user behaviour in tagging system done in [11] reports that the number of new unique tags decreases with the increase of the number of users. This suggests that users are using the same tags, which opens the possibility of pattern discovery through discovering how these tags are connected to each other. The same study [11] makes a categorization of tags into 9 categories: *identifying what (or who), identifying what it is, identifying who owns it, refining categories, identifying qualities, self reference* and *task organizing*. The first one includes the tags that are general in the sense that they represent general concepts known to most users, and the fourth one is used to better explain the tags that describe the high level objective category of the tagged resource. From such tag categories, some hierarchical structure between tags can be inferred. Therefore, some tags can be organized into hierarchical networks of association.

Another study [16] reports on the influences on the tagging behaviour of users. The authors studied an SBS by creating 4 experimental groups that were presented with no tags, all tags, popular tags and recommended tags, respectively. The authors found three major influences on tagging behaviour: user habits, community tagging, and the algorithm for selecting the tags that are presented to the user. Furthermore, the authors present a second categorization of tags into tag types: factual, subjective and personal. The study examines the utilization of tags and states that the major usage areas for tags are: self-expression, organizing, learning, finding and decision support. A major finding on the relationship between types and tasks is that while personal tags are most popular for organizing, factual tags are preferred for finding. From this brief survey of user behaviour in tagging systems, some conclusions for the task of a structure discovery algorithm can be drawn. First of all, the algorithm should show users only those tags that are general enough for everyone to understand their relation to the content of the resources “behind” the tag. This means that a content analysis of resources must be applied at some level. The other problem that should be solved is how the algorithm presents the tag structure to the user. This paper proposes to organize tags into hierarchical association networks.

2.3 Semantic and social web

Folksonomies depict users’ view of content and a different way of conceptualizing knowledge as opposed to expert-created taxonomies. Although some works [17, 10] suggest that folksonomies are by themselves enough to create an organization of knowledge, this does not mean that they are an equivalent of ontologies. Ontology as a model of knowledge is not the same as the taxonomic way of organizing concepts into hierarchical structure. In fact, the social web does not reject the hierarchical structure of knowledge as such, but rather the way it is built in traditional settings. If ontology is understood as “the specification of conceptualization” [5] then it should not be mistaken for one of the ways of its expression and design – centrally controlled categorization. Folksonomies reject centrally controlled knowledge structures, but centrally controlled categorization does have advantages when it comes to consistency, comprehensiveness, etc. Therefore, in order to create a system of collective intelligence it is necessary to extend SBSs by some other knowledge than the one which comes only from the users and their interaction with the system. In [6], the author suggests a model of tagging that incorporates a source into it, and represents it with 4 arguments as: *Tagging (object, tag, tagger, source).*

In this schema, s*ource* represents “objects universe quantification” [5] that corresponds to a namespace. This means that *sources* can represent different communities (applications) or different categories (resource *i* has been tagged with tag *c* by the user *a* for category *o*). This is a way of conceptualizing the ontology of folksonomy [6]. To make this connection between the subjective views of users and objective reality of resource content, in [7] the author suggests 3 different groups of methods: standard-based, information extraction and knowledge discovery.

The idea of semantic and social web combination through an SBS is investigated in a number of works. In [2], the authors combine tags into hierarchies using Word Net. The authors create a hierarchical network of tags enriched by the different type of relations between them. A combination of clustering and semantic enrichment is presented in [18], which uses Wikipedia relations between tags previously clustered into several groups. The authors first create a co-occurrence matrix of tags and then apply a clustering scheme to create groups of similar tags. After clustering, the Wikipedia topic hierarchy is used to discover the nature of relationships between tags in a cluster. These relationships could be simple IS-A relationships between tags or more complex relationships. An LSA (Latent Semantic Analysis) was applied to folksonomy data in [20] to find the latent connections between tags and then regard the latent factors as high-level ontological concept.

3 Topical structure discoveries

In this paper, we adopt the structure discovery approach to semantic and social web combination by association rules discovery. The standard Apriori algorithm [1] is modified to take into account the relevance of a resource to a specific topic. A topic *o* represents a tag that can be regarded as a concept with a wide-enough scope to describe an ontological category that is generally understandable in a domain in which a system functions.

First of all, it is necessary to identify topics. As shown in [20], the number of underlying topics in tagging systems is small.

3.1 Topic selection

In order to define topics we searched for tags that satisfy 3 conditions:

1. high frequency,
2. good descriptors of content,
3. existence at a high level of an expert-made hierarchy.

The motivation for setting the first condition has to do with topic definition. If we consider topics as terms that are well understood terms that identify some category that means that it used often by different users. In that way the subjective and personal tags are discarded and only factual are used as topic candidates. The second condition has the task of filtering factual tags into relevant and not relevant tags. For example, if a resource (a web-page) is tagged with tag *europe*, but is in fact a page describing Asia, that even if *europe* can be generally considered as a factual tag in this particular case it is irrelevant, so it has to be discarded. Finally, the topic has to understandable both to the users who tag the resources and to the users who use SBSs for searching. Therefore, in order to find the topics which all users can relate to we introduced the third condition. This condition has to check if experts have identified some concepts as general concepts which could be comprehended by all users of a system.

Condition 1 is easily checked with a simple count of occurrence of tags inside a folksonomy (as high, we consider the 20 most frequently used tags). The second condition means that a tag can be used to clearly identify content that is tagged with it. To test this clarity of identification, a classifier that considers tag as a positive class is built. The examples used to train the model are some resources that are not a part of a system with the same annotation. In our evaluation, we took the 200 first Google hits (using the tag as query) as positive examples, and 100 hits of several[1] semantically clearly different query words as negative examples. If this classifier had a high recall when applied to the resources tagged with the tag, we considered the tag as a candidate topic tag. For the third condition, we considered whether the tag (or its lemma or an obvious synonym) appears in one of the first five levels in an expert-created hierarchy. Concretely, we used Open Directory Project (dmoz.org) hierarchy. If a tag *c* satisfies all the conditions, then it is considered as representing the topic *o*.

3.2 Structure discovery

Once topics are identified, it is necessary to find the relevance of resources to the selected topics. The use of relevance or topic specificity is necessary to avoid the inclusion of subjective or deliberately false tagging. For example if we look at the set of three resources (song lyrics) which are tagged by users with three tags:

- *i1* tags: lyrics, love
- *i2* tags: lyrics, love
- *i3* tags: music, fun

If user searches for love song lyrics, it should be enough for him to query the system with tags *lyrics* and *love*. If we discover association rule from these tags, one of them will be *lyrics*➔*love (sp=0.66; cf=1)*. If confidence (*cf*) and support (*sp*) are calculated in this way (based on tags only), the results do not reflect the resource itself. However, the users do not search for tags, but for resources. “Love” can be used by some users to tag lyrics that are the lyrics of love song, and in that “mindset”, the tagging user uses the tag as a factual tag [16], but it could also be the personal tag of a user who reflected on lyrics he loves. In order to distinguish between these two, the relevance of a resource to a topic must be calculated. For example, let the relevance of i1 to a topic *love song* be 1, and relevance of resource i2 to the same topic be 0. Then the *lyrics*➔*love* association rule has different *sp* and *cf* values: 0.33 and 0.5 respectively. This means that for the query (*love* and *lyrics*), the probability of

[1] For our evaluation negative example query words were manually selected. This could be automated by using WordNet in the following way. If we sample a WordNet for a subgraph which contains term that is used as a positive class label, we can find semantically different terms by choosing the term that has the highest distance from the positive class label term.

a relevant resource being included in the answer set is 50%, which is more precise than the previous. For a given set *O* of topics *o*, the relevance of a resource $i \in I$ can be expressed for any topic *o* as *T(o,i)*. The value of *T(o,i)* can be determined in number of ways, by expert decision or ontology learning. In our algorithm, we consider only binary relevance obtained as a result of a classification model built to check for the second topic selection condition.

To find how the tags relate to each other we extend the association rule $x \rightarrow y$, where *x* and *y* are tags, to include topics. We express topic-specific association rule for a topic *o* and tags *x* and *y* as:

$$x \rightarrow_o y \tag{2}$$

To discover association rules ($x \rightarrow_o y$), we extend the well-known formalism that defines association rules by their support and confidence.[2] We introduce two new measures: relevant support (3) and relevant confidence for a chosen topic *o* and its representing tag *c* (4):

$$suppR(o,c,x,y) = \frac{\sum_{k=1}^{d} T(o,i_k) \times |\{a | a \in A \wedge (a,i_k,c) \in F \wedge (a,i_k,x) \in F \wedge (a,i_k,y) \in F\}|}{\sum_{k=1}^{d} |\{a | \exists z \in C : (a,i_k,z) \in F\}|} \tag{3}$$

$$confR(o,c,x,y) = \frac{\sum_{k=1}^{d} T(o,i_k) \times |\{a | a \in A \wedge (a,i_k,c) \in F \wedge (a,i_k,x) \in F \wedge (a,i_k,y) \in F\}|}{\sum_{k=1}^{d} T(o,i_k) \times |\{a | a \in A \wedge (a,i_k,c) \in F \wedge (a,i_k,x) \in F\}|} \tag{4}$$

Using these two measures, association rules are discovered, and then based on these rules a network of association is created using the approach described in [14] and [15]

4 Case study and evaluation

In this section, we describe the results of a first evaluation of the method. This case study investigated tags from bibsonomy and compared them to the category structure from the Open Directory dmoz.org. All tests were run on a set of data from bibsonomy (www.bibsonomy.org) from 30/06/2007. The data used was only from the bookmarking part of the system (the system also allows for scientific publication tagging). It consists of 78544 resources with 163298 tag applications of 19978 distinct tags by 901 users.

The method consists of 4 parts:

- choice of topic(s);
- association rule discovery;
- creation of the network of association;
- evaluation;

[2] In traditional association rule mining, the support of $x \rightarrow y$ is the number of transactions containing x and y divided by the number of all transactions. The confidence of $x \rightarrow y$ is the number of transactions containing x and y divided by the number of transactions containing x [1]. Here, "tagging actions" (user-item relations) are the transactions, a transaction contains a tag if this tag was used by this user for this item, and filters ensure that the rule is relevant in the context of the topic and its representing tag. These adaptations are expressed in (3), (4).

As stated in section 3, we set three requirements that a given tag must meet in order to be considered as a topic. Based on the first one (tag frequency), we chose 4 topic candidates to run our tests. For the case study, we chose the following 4 tags: *web2.0*, *linux*, *programming* and *semanticweb.* The first tag (*web2.0*) was chosen because of its highest frequency, and the others were randomly selected from the tags that were in English. Table 1 shows the top 20 tags and their frequency in the used data set.

Table 1. Frequency of top 20 tags in the dataset (highlighted tags are considered topic candidates in the case study)

	Frequency	tag		Frequency	Tag
1	1904	web2.0	11	955	java
2	1675	allgemein	12	930	tagging
3	1646	blog	13	907	search
4	1593	software	14	764	research
5	1563	tools	15	764	semanticweb
6	1396	linux	16	763	news
7	1265	web	17	750	opensource
8	1206	reference	18	744	design
9	1174	programming	19	740	webdesign
1	1144	internet	20	722	wiki

The second condition called for tags that are good descriptors of content. To address this, a classification model was trained. In lack of a generally accepted truth we relied on Google for providing us with examples. For each of candidate tag (*web2.0*, *programming, semanticweb, Linux)*, the positive examples were the 200 (+/- due to the fact that some pages could not be retrieved) web pages on English language that are returned by Google when one of the tags is used as a query. As negative examples, we used 100 web pages (5 * the first twenty hits) which were given by Google as results for the following 5 queries: gardening, usability, cartoon, graphic, design. Using these inputs (200 positive and 100 negative examples), an SVM document classifier is trained for each candidate tag, using the Weka 3.5 machine learning tool. Kernels and parameters for each classifier were chosen to minimize the error, using 10-fold cross validation as the evaluation setting. Recall and precision for best performance classifiers for each of 4 tags is shown in table 2.

Table 2. Precision and recall of the classifiers learned from Google hits.

	web2.0	*Semanticweb*	*programming*	*linux*
precision	0.68	0.89	0.78	0.95
recall	0.75	0.82	0.85	0.93

After this, we used the trained models to assess the description power of a tag by measuring recall of the resources tagged with a candidate tag and 100 randomly selected resources that are not tagged with the candidate date set. In Table 3, we show the recall values from that result set. Since the classifiers model the ‘ground truth’, the recall shows how well users used a tag as a concept name for the ‘ground truth’ model. It also shows whether the resources can be classified using the tag name as a discriminative function, and it indicates a level of user ‘agreement’ on the concept that could be expressed by a keyword (tag). Table 3 shows the results for the 4 candidate tags.

Table 3. Recall results for built models with the respect to the ‘ground truth’ models

	web2.0	*semanticweb*	*programming*	*Linux*
recall	0.65	0.91	0.96	0.97

For one tag to be considered as good descriptor of content its recall must be over a predefined threshold. We set this to 85%, a value that should be evaluated against human judgments in future work. Due to its low recall, the tag *web2.0* is discarded as a candidate topic set. The low score could be explained as a result of very broad meaning of a term Web 2.0 and a number of themes ranging from business, entertainment, IT and other areas that can be described by this tag.

To satisfy the third condition we searched the Open Directory hierarchy to find top level entries that correspond to the candidate tags. Minor spelling differences were disregarded. This could be automated in a straightforward way. For the tag *programming*, an entry was found on the second (not considering the top level named TOP) hierarchical level (http://www.dmoz.org/Computers/Programming/), for the tag *linux* an entry was found on the fourth level (http://www.dmoz.org/Computers/Software/Operating_Systems/Linux/), and for the tag *semanticweb* an entry was also found on the fourth level (http://www.dmoz.org/Reference/Knowledge_Management/Knowledge_Representation/Semantic_Web/). Although all candidates have a corresponding tag in the directory, the last one (*semanticweb*) does not have any child entries. However, the method for finding hierarchical structure presented above finds hierarchical structure below the topic. Therefore, the method would produce a network which is incomparable with the structure of this keyword in dmoz.org (no nodes below it).

The next step in our method was to create association rules and to use them to derive hierarchical networks. This was done in order to address goals 2 and 3. For each candidate tag, we define two data sets: the *whole corpus (w)* dataset that represents the topical subset for a candidate tag (as defined in equation (4) above) and the *relevant corpus (r)* dataset which includes only those resources that have been classified by the classification model as positive classes (e.g., for *programming*, there are 1174 resources in *w* and *959* in r, for *linux* 1396 in *w* and 1354 in *r*, for *web2.0* 1904 in *w* and 1275 in *r*, for *semanticweb* 764 in *w* and 695 in *r*). After this, for both corpora association rules are learned, and an association network is created using the method described in [15], see Section 2.1.

In order to compare these networks with an expert-created network, a network was derived from the Open Directory Project in a following way. For each tag that appears in the generated association network, we searched for the corresponding concepts inside the dmoz.org category network and included a complete path (all the nodes) from the category to the concept that was found. For example, if we are searching for the *javascript* in a *programming* data set we add it to the network only if it exists in dmoz.org under the *programming* category, and also add all the categories from programming to javascript (Programing:Languages:JavaScript). The derivation of such network and not the use of the entire dmoz.org category network are necessary to overcome differences in the number of concepts (tags), since bibsonomy.org is a research project with a limited number of tags. Figure 1 shows the created network of associations for the whole *programming* corpus. Figure 2 shows the created network of associations for the relevant corpus of tag *programming.* Figure 3 shows the network derived from dmoz.org for *programming.*

Once all 3 networks were created, we calculated the similarity between them using the taxonomic overlap measure described in [12]. This measure calculates the similarity of taxonomic networks by calculating the overlap of two nodes' respective set of parents and children. Table 4 shows the taxonomic overlap between *programming* and *linux* whole and relevant networks and dmoz.org derived.

Table 4. Taxonomic overlap for 3 derived networks for the tags *programming* and *linux*

	Programming			Linux		
	whole	relevant	dmoz.org	whole	relevant	dmoz.org
whol		0.5982	0.3664		1	0.7923
relev	0.8059		0.7033	1		0.7923

The results in table 3 (for *programming*) show that the similarity between dmoz.org structure and structure derived using the approach described in the paper. The results in table 3 (for *linux*) show no improvement with the topic inclusion which can be explained by the high precision of tag *linux* – out of 1122 documents which were tagged with linux that could be retrieved, 1105 were classified as relevant, and both *w* and *r* corpora produced the same networks of association.

The high TO similarity between the dmoz.org structure and relevant corpus structure for both tags implies that there is a high similarity between expert-generated structure and structure that emerged from user interaction with the system.

This evaluation is only a first step in a thorough assessment of our method. Its main aims were to validate the basic approach by testing whether a random test set would produce "good-enough" quantitative results, and to gain some qualitative insights into the compared structures. For instance, the results showed that a pre-selection of domains may be needed to focus on topics that are "basic-level concepts" worthy of further subdivision both for experts and for taggers (unlike the example *semanticweb*). They also showed that while the expert-generated hierarchies are often more semantically constrained than those created by taggers (for example, *ajax* is a specialization of *javascript* in Fig. 3, whereas in Fig. 2 its frequent occurrence and central importance to users active in the bookmarking platform lift it to the level directly below *programming,* and make *web2.0* a descendant – essentially showing that tag co-occurrence not always indicates an IS-A relationship, but may indicate some other strong association). In future work, this will be extended by a more comprehensive quantitative evaluation that also compares this approach to other ways of deriving tag hierarchies such as the ones described in [10, 14].

5 Conclusion

Research presented in this paper showed an approach to structure discovery in folksonomies by combination of tag structure and content analysis. The goal was to show and evaluate a possible way of interaction between social and semantic views of web content organization.

The motivation of the paper was to show the use of data mining in an attempt to combine social and semantic web into a single framework. By evaluating three goals, we showed that users can specify categories using certain tags as well as state-of-the-art search engine, and that based on a combination of user inputs, we can artificially create such structures that are similar to structures created by human experts. We also showed that these structures could be enhanced by applying content analysis.

Future work should go more into direction of usability and try to measure how usable different structures are that could be discovered inside folksonomies, and whether topic enrichment can produce better results. Besides not tackling the usability side of the approach, our algorithm has some limitations. First, by setting the first condition of topic choice, we disregard those tags that are not frequent, and this could result in a failure to recognize such topics that are well structured in folksonomy by a limited number of users and resources and therefore a smaller number of tag applications. In such a way, small communities would be omitted from topic discovery. A possible way to overcome this is to use a different selection criterion for the first condition such as weights similar to tf.idf weight (used in information retrieval) applied to tags. Since our goal was not to build a better text classifier, we used a simple pre-processing methods (stemming) and a well-known text classifier method. The effects of choosing different pre-processing and mining methods could be investigated in further work. Also the algorithm applies only to the textual resources (documents, HTML pages), but many social bookmarking engines allow users to tag multimedia documents, and this is not considered in this work, although a similar approach could be applied to multimedia using a different kind of ‘ground truth’ (for example, given by image annotations). In order to completely evaluate our algorithm and the results, a study on a larger and more commercially oriented social bookmarking engine would be useful.

In sum, many of the approaches that emphasize collective intelligence as a fundamental element of Web 2.0 disregard the statistical (data mining) element that is in fact the underlying element of collective intelligence. As Web 2.0 brought a social revolution to the internet, we believe that a rise of Web 3.0 applications (www.twine.com, www.opencalais.com) and research results will bring some order into the chaos that came with Web 2.0.

References

1. Agrawal, R.; Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB; Bocca, J. B.; Jarke, M.; Zaniolo, C., Eds.; Morgan Kaufmann: 1994.
2. Angeletou, S., Sabou, M., Specia, L., Motta, E., (2007) Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report. Workshop: Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference.
3. Atlee, T., and Pór, G. (2007) Collective Intelligence as a Field of Multi-disciplinary Study and Practic,, http://www.community-intelligence.com/blogs/public/2007/01/a_source_document_for_collecti.htmlm, retrieved on 18/10/2007
4. Begelman, G., Keller, P., and Smadja, F. (2006) Automated Tag Clustering: Improving search and exploration in the tag space. In Proceedings of the Fifteenth International World Wide Web Conference (WWW2006) (Edinburgh, Scotland, May 22-26, 2006). ACM Press, New York, NY, 2006.
5. Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud., 43(5-6):907-928.
6. Gruber, T. (2007). Ontology of Folksonomy: A Mash-up of Apples and Oranges, Published in Int’l Journal on Semantic Web & Information Systems, 3(2), 2007. (Originally published to the web in 2005)
7. Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web, 6(1):4-13.
8. Hammond, T., Hannay, T., Lind, B., and Scott, J. (2005), Social Bookmarking Tools (I), D-Lib magazine, vol. 11, no. 4, 2005, p. 1082
9. Hassan-Montero, Y., and Herrero-Solana, V. (2006) Improving Tag-Clouds as Visual Information Retrieval Interfaces. In Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies (InSciT ‘06) October 25-28, 2006, Mérida, Spain.
10. Heymann, P., and Garcia-Molina, H., (2006), Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Stanford InfoLab Technical Report 2006-10.
11. Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. J. Inf. Sci., 32(2):198-208.
12. Maedche, A., and Staab, S.: Comparing Ontologies - Similarity Measures and a Comparison Study. Internal Report 408, Institute AIFB, University of Karlsruhe, 2001.
13. Mika, P. (2005). Ontologies Are Us: A Unified Model of Social Networks and Semantics, Web Semantics Science Services and Agents on the World Wide Web (ISSN: 1570-8268), vol. 5, no. 1, 2007, p. 5.
14. Schmitz, C., Hotho, A., Jaschke, R., and Stumme, G. (2006).Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. iberna, editors, Data Science and Classifcation, Studies in Classifcation, Data Analysis, and Knowledge Organization, Berlin, Heidelberg, Springer, pages 261-270.
15. Schwarzkopf, E., Heckmann, D., Dengler, D., and Krner, A. (2007). Mining the structure of tag spaces for user modeling. In Complete On-Line Proceedings of the

Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling, pages 63-75, Corfu, Greece.

16. Sen, S., S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, M. F. Harper, and J. Riedl (2006). Tagging, communities, vocabulary, evolution. In CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, New York, NY, USA, pp. 181-190. ACM Pres
17. Shirky, C. (2006), Ontology is Overrated: Categories, Links, and Tags, http://www.shirky.com/writings/ontology_overrated.html, retrieved on 16/11/2007
18. Specia, L. and Motta, E., (2007) Integrating Folksonomies with the Semantic Web. In Proc. of ESWC'07, 2007.
19. Van der Wal T. (2004), Would We Create Hierarchies in a Computing Age?, December, 17. 2004 http://vanderwal.net/random/entrysel.php?blog=1598, retrieved on 16/06/2008
20. Wu, H., Zubair, M. , and Maly, K. (2006), Harvesting social knowledge from folksonomies, Proceedings of the seventeenth conference on Hypertext and hypermedia, August 22-25, 2006, Odense, Denmark
21. Wu, X., Zhang, L., Yu, Y.(2006). Exploring social annotations for the semantic web, Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland

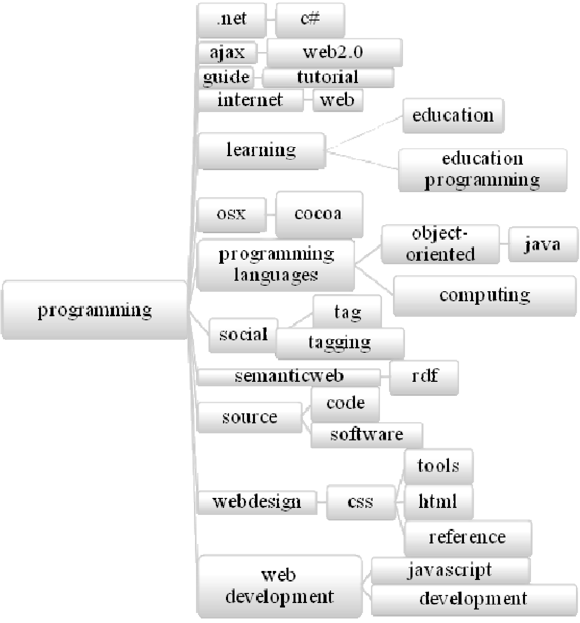

Fig. 1. Network of associations for the whole *programming* corpus.

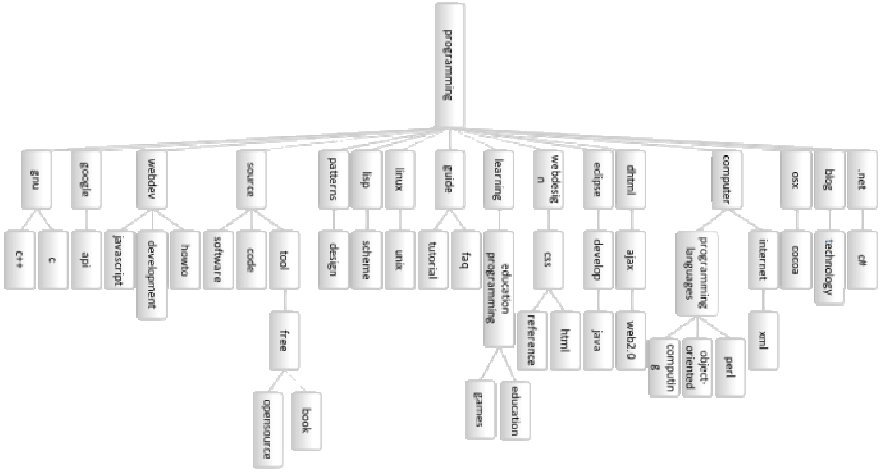

Fig. 2. Network of associations for the relevant *programming* corpus.

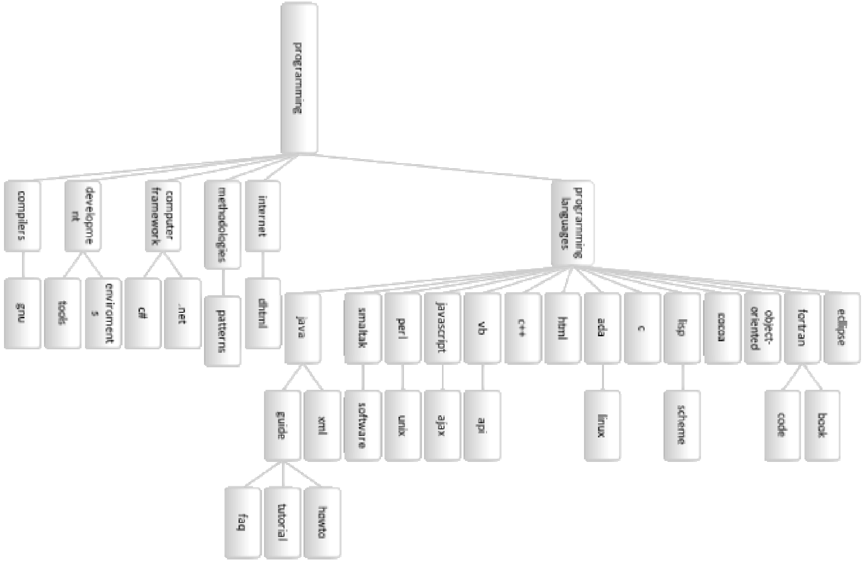

Fig. 3. Derived dmoz.org network of associations for *programming* corpus.