System development enrichment authority file records with social annotations.

DAGLAS SPYROS

SUMMARY OF THE DIPLOMATIC THESIS

UNIVERSITY OF ATHENS

MSc Information Systems

The mentioned thesis was supervised by Associate Professor Division of Archives and Libraries of Ionian University, Christos Papatheodorou and candidates of the Department of Archives and Libraries of Ionian University, Dina Kakalis. Our thesis was done in collaboration with graduate students in the Department's information systems and Business, Koumakis Marina Kakavouli Dionysius. Subject of our thesis was to develop an authority file records enrichment system with social annotations. The increasing access to social networking services (such as Facebook, LinkedIn, etc.), and the development of Web 2.0, helped for "crossing" the borders and the integration of people into a new, digital landscape. As the world moves ever closer to embedded technology in traditional social attitudes, there is greater need for innovative research and development in various aspects of the social aspect of computing.
In this thesis, we dealt with finding methods of enrichment with social annotations, files, standard terms. The social annotations, appeared with the first application services Web 2.0. The term Web 2.0 (Web 2.0) is used to describe the new generation of Web, based on the increasing ability of Internet users to share information and collaborate online. This new generation is a powerful online platform where users can interact without any special knowledge on computers and networks.

Social networks are an approach for investigating social structures. More specifically it is a group of collaborating individuals and / or competing individuals or even companies that are related with [2].

The philosophy of library services is user-centered, which existed before the development of Web 2.0. However, with the advent of social interaction technologies such as blogs and wikis, libraries have seen a radical change. Libraries recognize the fact that the technologies of Web 2.0, can provide users with many services, and implement the 'social software' in innovative ways. The technologies of 'social software' found uses in libraries, created the need for a new term, the Library 2.0.

Below, we will do a brief overview of the problem we face and the solution we approach. More specifically, our work focuses on developing a social recommendation tool with annotations to enrich the thematic description of the records for libraries. Data of our work were:

1. The record established records (authority file).
2. A social taxonomy (folksonomy), which consists of social annotations of users and comes from LibraryThing,
3. The log file of queries in the library catalog to search for resources-documents (searches based on the following indexes: subject, author, title, year of issue, ISBN, language, notes, publisher, title number and other fields).

Starting our investigation, we examined the percentage of coverage - in lexicographical level - between the main sets of our data, as mentioned previously. Specifically, we sought the coverage of the following sets:

- Value F ∩ A: (Social Archive ∩ taxonomy established records - Folksonomy ∩ Authorityfile)
- Value A ∩ Q: (Archive records established ∩ log queries - AuthorityFile ∩ Queries)
- Value F ∩ Q: (Social taxonomy ∩ log query-Folksonomy ∩ Queries).

The results led us to conclude that we must find a way to bridge the gap between the language that the users use, and the annotations, in order to represent more users and to yield better results in the search for evidence.

The standard terms file of the Panteion University of Athens, which we received for our research, contained data that we needed throughout our research. Should, therefore, to process

the file and create a form that suited our purposes and voithage in our work. After the necessary treatment of established archive records, we were in the form:

Παιδαγωγική Ιστορία
Μνήμη--Κοινωνικές απόψεις--Ευρώπη
Memory--Socialaspects--Europe—
Εγκληματικότητα ανηλίκων--Ελλάδα
Ιστορία
Juvenile delinquency----Greece—
History
Προλεταριάτο, Δικτατορίατου
Δικτατορία του προλεταριάτου
Κομμουνιστικό κράτος

After finding the overlap of these sets and editing the file of standard conditions, we examined the relationship of the annotations we have in our possession, with the access points for our records. More specifically, we noticed what annotations, there are such, as conditions in the subjects, the author, title or any other field of a record. Such an annotation, we do not take in mind for the next stages of our research, we simply say that covers a user's need, to manage in this way the information. Thus, we seperate what annotations were useful for us and we were working on the continuation of our research.

For the annotation we keep, we distinguish two possible cases:

- The apostille is a random personal annotation from a user and is not important to all of our users or even for information recording.
- The annotation is important and the possibility is discussed in the next step, this evaluation of the annotation.

The evaluation of an annotation is divided into two stages. The first is the extraction from the LibraryThing and Google Books informations, that will give us any annotation, while the second stage is the combination of the results obtained and the conclusion we did for them.

The value of the annotations that we consider useful for our system is a combination of information we receive from the LibraryThing and Google Books. In fact, in LibraryThing, will see the number of impressions that each annotation is seen talking to a presumption. While in Google Books, we will see if this annotation is important to the presumption, as we look at the number of pages displayed in it.

With this information we receive, we will evaluate each annotation, not only according to the preference of the users in it for a specific item, but also be counted and the importance of annotation to be presumed, according to the writings of each author.

The application that we designed for extracting information from LibraryThing, aims to give us for each item - book we want, all annotations that users have been added, and the number of performances. The site LibraryThing, contains information to a DOM tree. We managed and selected the appropriate branch of the tree and get the necessary information. The results you get for each item was:

110037373	Losing control? : Sovereignty in an age of globalization	Box	17
110037373	Losing control? : Sovereignty in an age of globalization	Economics	1
110037373	Losing control? : Sovereignty in an age of globalization	economy	1
110037373	Losing control? : Sovereignty in an age of globalization	Globalization	3

In the above table, we see some results, as we received from LibraryThing. The first column is the rec_id of the presumption. The rec_id, is a unique number for each item, which we had given the library of the Panteion University. This number was the fundamental key to our tables, to process and combine our results from LibraryThing and Google Books. In the second column is the title of the presumption that we seek, while the third and fourth are an annotation that has been used for the item and the number of users, respectively.

In search of evidence, we found ourselves faced with several. Problems such as the mismatch between the isbn of a book that had the opacial and similar to that has been designated to LibraryThing, or recognition from our server site as malicious program (bot), resolved after persistent efforts and tests for optimal solution on the time of execution of our program and the quality of data mining.

The extraction of information from GoogleBooks, is similar to the procedure for LibraryThing. For more information we recommend the dissertations of graduate students Koumakis Marina and Kakavoulis Dionysius.

After the above processes, we got 41470 annotations of 4018 records. In the following section, we complete our research, with experiments performed by us to complete our system.

The first step in the implementation of our experiments was the calculation of the shape weighting procedures tf-idf. The tf-idf measure of weight, often used in information retrieval. The measure estimate how important a term proportional increases the number of the term appears in the document, but is offset by the frequency of collection. So, we calculated the shape of tf-idf, for annotations, separately for those who got the LibraryThing and from Google Books. Then, we created a table that combined the above results. More specifically, we calculated a value of tf-idf, which is derived from both individual values described above, a certain weight at a time, ie, $tfidf = (a * tfidf_{LibraryThing}) + (b * tfidf_{GoogleBooks})$. The values we have in a and b, is from 0% to 100%, opposite each other. Below, we present the table created for some of the annotations that we calculated.

REC ID	TAG	GB tfidf	LT tfidf	100%GB-	90%GB -	80%-20%	70%-30%	60%-40%	50%-50%	40%-60%	30%-70%	20%-80%	10%-90%	0%-100%
10011095	textbook - sacred texts	0	1,69689	0	0,16969	0,33938	0,50907	0,67876	0,84845	1,01814	1,18782	1,35751	1,5272	1,69689
10037815	advocacy*	7,386	7,38585	7,386	7,38599	7,38597	7,38596	7,38594	7,38593	7,38591	7,3859	7,38588	7,38587	7,38585
10035721	advocacy*	7,386	1,84646	7,386	6,83205	6,27809	5,72414	5,17019	4,61623	4,06228	3,50832	2,95437	2,40042	1,84646
10048574	gender inequality	0	1,94783	0	0,19478	0,38957	0,58435	0,77913	0,97391	1,1687	1,36348	1,55826	1,75305	1,94783
10013628	gender inequality	0	1,94783	0	0,19478	0,38957	0,58435	0,77913	0,97391	1,1687	1,36348	1,55826	1,75305	1,94783
10061635	dystopia	0	1,55826	0	0,15583	0,31165	0,46748	0,62331	0,77913	0,93496	1,09078	1,24661	1,40244	1,55826
10064230	Customer Service / Loyalty	0	0	0	0	0	0	0	0	0	0	0	0	0
10055365	Sandra*	0	4,24223	0	0,42422	0,84845	1,27267	1,69689	2,12112	2,54534	2,96956	3,39379	3,81801	4,24223
10056885	philosophy of mind*	0	0	0	0	0	0	0	0	0	0	0	0	0
10058229	philosophy of mind*	0	1,52164	0	0,15216	0,30433	0,45649	0,60866	0,76082	0,91299	1,06515	1,21731	1,36948	1,52164
10019919	philosophy of mind*	0	0	0	0	0	0	0	0	0	0	0	0	0
10032365	philosophy of mind*	0,11565	0	0,11565	0,10409	0,09252	0,08096	0,06939	0,05783	0,04626	0,0347	0,02313	0,01157	0
10060795	philosophy of mind*	0	0,14579	0	0,01458	0,02916	0,04374	0,05831	0,07289	0,08747	0,10205	0,11663	0,13121	0,14579
10043741	philosophy of mind*	0,42609	0	0,42609	0,38348	0,34087	0,29826	0,25565	0,21305	0,17044	0,12783	0,08522	0,04261	0
10046145	philosophy of mind*	0	2,02886	0	0,20289	0,40577	0,60866	0,81154	1,01443	1,21731	1,4202	1,62308	1,82597	2,02886
10043431	philosophy of mind*	0,18261	3,04328	0,18261	0,46868	0,75474	1,04081	1,32688	1,61295	1,89901	2,18508	2,47115	2,75722	3,04328
10043436	philosophy of mind*	0,48696	0	0,48696	0,43826	0,38957	0,34087	0,29218	0,24348	0,19478	0,14609	0,09739	0,0487	0
10062504	disabilities*	0,11878	2,82815	0,11878	0,38971	0,66065	0,93159	1,20253	1,47347	1,7444	2,01534	2,28628	2,55722	2,82815
10046346	Staatsrecht*	2,79972	8,48446	2,79972	3,36819	3,93667	4,50514	5,07362	5,64209	6,21057	6,77904	7,34751	7,91599	8,48446

Table 1. $tfidf = (a * tfidf_{LibraryThing}) + (b * tfidf_{GoogleBooks})$.

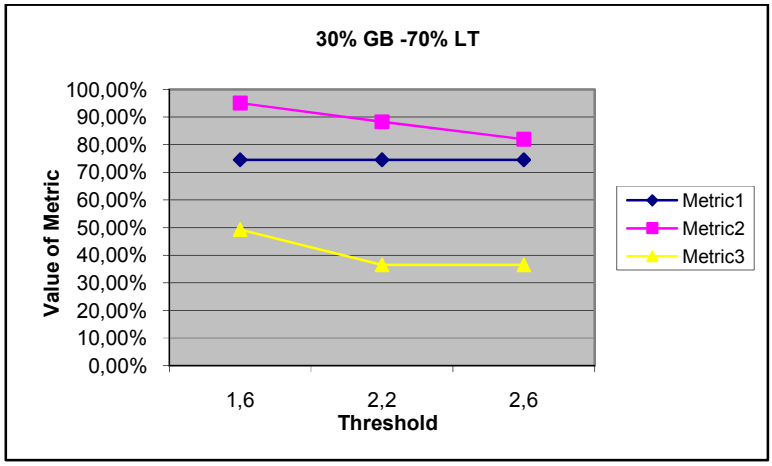
To continue our research, we chose the results of columns from 70% GB-30% LT, till 30% GB-70% LT, in which we applied 3 different metrics that we had alternatively adding annotations to the file contents of standard conditions. The columns selected in the figures, we felt that we represent the choice of most librarians, but does not exclude a different choice. Initially we thought 3 different thresholds (threshold) in the total price tfidf - for each combination of a and b, as mentioned above - with values 1.6, 2.2, 2.6.

For each threshold we calculated the following indicators:

- In our first experiment, we wanted to see how the increased coverage of the entire query log, the whole authority file records.
- In the second experiment, we examined the percentage of items for which enriched our knowledge. We saw how many items from all the 4018 items we had, we managed to increase the enrollment in terms of established archive.
- In our third and final experiment, knowing all the annotations that extraction from the LibraryThing and Google Books, for a particular item, we noticed how many of these annotations were added in the end, to file standard terms.

For the second measure, we observed that for a rate close to 90% of all the evidence we had at our disposal, increased entries in the file of standard conditions. This percentage is very high and demonstrates the value of our research, while shows the accuracy of the methods followed. Certainly, the reduced threshold, the more useful and annotations are added to file standard terms, so the more the evidence is concerned. However, one should bear in mind, that the proper choice of threshold is a very important factor in assessing the results we present as a small backyard means that the file added more annotations but with less respect and value for that item. In contrast, a higher threshold means that the file is added fewer annotations and more than value for a specific item, but may lose some useful annotations on the record. More information about the other two measures performed an experiment of our research can be found in diplomatic work of Marina Koumakis (Measure 3) and Kakavouli Dionisius (Measure 1).

Our results for a rate 70% LibraryThing and 30% Google Books shown below:



In the field of social networks and the analysis thereof, as well as the "world" of Library 2.0, there are some open issues that need investigation and attention by the entire research community. There are concerns where professionals of libraries show and should be considered. These concerns relate to issues related to trust the accuracy, responsiveness, reliability and ethics on the use of applications of Web 2.0.