Magnifico: A Platform For Expert Mining Using Metadata

Na Li, Lei Zhou, Denis Gillet
École Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{na.li, lei.zhou, denis.gillet}@epfl.ch

ABSTRACT
In this paper, a modified TF-IDF approach is proposed to recommend experts by inferring search query topics using metadata. A multi-disciplinary reputation metric is also introduced to select people with specific expertise. The approach is validated on Mendeley corpus using a prototypal expert mining platform, *Magnifico*. The user interface of the platform is also discussed.

Author Keywords
Information Retrieval; Reputation Systems; Ranking.

ACM Classification Keywords
H.4.0 Information Systems Applications: General; H.3.3 Information Search and Retrieval: Information filtering.

General Terms
Human Factors; Design.

INTRODUCTION

Reputation systems in online communities typically adopt global reputation metrics, where for each entity, an overall reputation score is computed to reflect the general trustworthiness of this specific entity. However, in many application scenarios such as personal learning environments (PLEs) [1], personnel recruitment, and conference program preparation, there is a clear need to find people that have particular subject-matter expertise. To comply with the requirements of multi-disciplinary environments, context-dependent reputation mechanisms should be designed to measure the expertise of people in diverse disciplines. Motivated by this, we modified the TF-IDF approach and designed a search application, namely *Magnifico*, recommending people with specific expertise based on user's query.

The *Magnifico* platform takes user's query as input, infers the possible disciplines of the query, and generates a list of people having the expertise in the disciplines, sorted by their expertise ranking. For each person in the ranked result, her publications and profile information is presented, as well as an expertise cloud indicating her major strengths. A *Magnifico* score is also shown to reveal how proficient she is in the particular discipline. Moreover, *Magnifico* supports filtering within the search result based on different criteria including people's academic status and research disciplines. It is worth mentioning that *Magnifico* provides suggestions of fields that the query may fall into. We believe that the suggestions can help users to progressively adjust and refine their search queries to obtain better results. The search approach we used and the system interface will be addressed hereafter.

DATA PREPARATION

The dataset used in *Magnifico* platform is the Mendeley corpus [2] provided in the framework of the HCIR Challenge 2012. It contains 1 million person profiles and 0.1 million academic publications with associated metadata including reader statistics, title, publisher name, and so on.

To enrich the dataset and provide better search results, we crawled additional metadata using Mendeley API [3]. The crawled metadata includes academic status of people, profile URLs, profile photos, and sub-disciplines within each main discipline in the given corpus. Furthermore, to filter out the non-English entries in the dataset, Compact Language Detector library [4] provided by Google Chrome was used to detect the languages of the metadata.

Within the Mendeley corpus, every publication has at least one author with a registered Mendeley profile. The other authors do not necessarily have profiles in the dataset. To solve the identify resolution problem, we used the first name, last name and the research discipline to match the unidentified authors with their profiles. For some people, their names in the profiles are inconsistent with what they are referred to in the author list of publications. For instance, the first name "Peter" in a profile could be referred to as "P." in the author list of a publication. Therefore, name abbreviations were also taken into account in the data matching process. If multiple profile matches were found for one author, the profile having the same research discipline as the identified co-authors was identified as the real one.

THE APPROACH

To find the people having the particular expertise that matches the search query, two problems should be solved: to infer what disciplines the search query falls into, and to select people with competence in those disciplines. Our solutions to the two problems are addressed in this section.

TF-IDF

Inspired by the Term Frequency – Inverse Document Frequency method (TF-IDF) [5], we propose a modified Term Frequency – Inverse Discipline Frequency approach to determine how relevant a given query is to a specific discipline. The original TF-IDF measures how important a word is to a document in a given corpus. The term frequency (TF) value is calculated by counting the number of times a given term occurs in a document. To avoid a bias towards long documents, this value is usually divided by the maximum term count of any word in the document. The TF value is normally used as a measure of the importance of a term within a specific document.

The inverse document frequency (IDF) value is used to determine whether a term is common or rare across all the documents in a given corpus. Dividing the total number of documents by the number of documents containing a term and taking the logarithm of the previous result, we get the IDF value of the term. The product of TF and IDF is often used to filter out the common terms such as articles and prepositions. Words that are common only in a small group of documents tend to have a higher product of TF and IDF than the ones that are common across all the documents.

Modified TF-IDF

Instead of determining the importance of a word within a document, we measure the importance of a word for a given discipline. To achieve that, we make use of the metadata associated with the profiles and publications in the given Mendeley corpus. For each profile, the metadata contains the main research discipline of that person, and for each publication, the metadata includes title, publisher name, the profile ID of one author, and the distribution of readers by discipline. Compared to the main research discipline (e.g., Computer and Information Science) of a profile, the disciplines of readers are given in more fine-grained categorizations (referred to as sub-disciplines in the rest of the paper) such as "Information Retrieval", "Artificial Intelligence", and "Computer Security". A main research discipline is composed of a few sub-disciplines.

For each publication, although the discipline categorization is not provided in the dataset, we believe that the main research discipline of the authors is a good indicator of how the publication can be roughly categorized. Additionally the sub-disciplines of the readers suggest more specifically the possible topics that the publication could fall into. The proportion of readers from a specific sub-discipline could be seen as the probability of the publication falling into that sub-discipline. For instance, if a publication has five readers from the field of "Information Retrieval" and ten readers from the field of "Artificial Intelligence", it suggests that the publication addresses the topic of "Information Retrieval" and "Artificial Intelligence" with the probability of 1/3 and 2/3 respectively. It could happen that a paper addressing a Chemistry issue has a few readers from Biological Sciences. To filter out the data noise, only readers from the same research discipline as the authors are taken into account.

As stated previously, for each publication, a probability vector is generated to reveal the distribution of possible sub-disciplines that publication falls into. We then make the common assumption that a publication is "a bag of words" [6]. Therefore, the discipline probability vector of a publication could be seen as that of the words appearing in that publication.

Afterwards we measure the importance of each word for a given sub-discipline using the term frequency of that word occurring in the specific sub-discipline. For every publication, after collecting all the words appearing in the title and publisher name, stop words are removed from the word collection. The stop words list used is the standard set of English stop words [7], extended with a few other common terms including "journal", "conference", "ieee", and so on. The "clean" collection of words will be used for term frequency calculation. For each word in the collection, we increase its term frequency (TF) value in all sub-disciplines of the particular publication, weighted by the probability of the corresponding sub-discipline we get from the probability vector.

After going through all the publications, a probability matrix is generated, revealing the term frequency of every word in every sub-discipline. However, the numbers of publications vary in different sub-disciplines. For instance, there are a lot more publications in the field of Biotechnology than that of Music. Due to that, there is a bias in the term frequency matrix towards the sub-disciplines having a large number of publications. For the purpose of normalization, we then divide the term frequency value by the total number of publications in the specific sub-discipline. In the end, the normalized matrix represents the sub-discipline distribution for every word.

Although the stop words have already been removed from the word collection, there are still some common words across all the sub-disciplines, such as "analysis", "research", "study", and so on. Those words are not discipline-specific and might create noises when inferring the sub-disciplines of the search query. To detect those words, a modified IDF value is computed as a measure of whether a word is common or rare across all the sub-disciplines. Unlike computing the original IDF value, we divide the total number of sub-disciplines by the number of sub-disciplines containing a term and take the logarithm of the previous result. Words with a high modified TF value but a low modified IDF value are considered as not discipline-specific, and are filtered out to reduce noises.

Finally, to infer what sub-disciplines a search query falls into, a sum is computed for the sub-discipline distribution vector of every word in the query. The top sub-discipline with the highest probability value is considered as the dominant topic of the query. If there are more than one sub-

discipline having a close probability value to the highest value, they are all seen as the dominant topics of the query, since there could be cross-discipline queries.

Multi-disciplinary Reputation Metric

After inferring the dominant topics of the search query, the next issue to be tackled is to select people with specific expertise from the collection of profiles and authors. In this section, the approach for computing people's expertise scores is discussed.

As the citation information of publications is not given in the Mendeley corpus, the reader count is used as a measure of how popular a publication is. For each publication, the readers who are not from the same research discipline as the authors are again filtered out. We then increase the authors' expertise score in a particular sub-discipline by the reader count in that sub-discipline. At the end of the iterations throughout all the publications, a matrix is produced, indicating the expertise scores (referred to as *Magnifico* score in our platform) of all the authors in the corresponding sub-disciplines.

Using the dominant sub-disciplines of the search query, a list of people with the corresponding expertise is obtained, sorted by their *Magnifico* scores. It is worth noting that, if a search query falls into more than one dominant fields, the search results could be either people having at least one expertise, or those having all the expertise at the same time, depending on users' choice. We believe that both cases could be useful in different application scenarios. For instance, a job recruiter could be interested in finding job candidates with the competence of either "Econometrics" or "Economic Systems". But an organizer of an e-learning conference might look for people who are proficient in both "Educational Technology" and "Information Science".

THE INTERFACE

The *Magnifico* platform is developed using Bootstrap [8] and Ruby on Rails [9] as frontend and backend frameworks. The user interface is illustrated in Fig. 1.

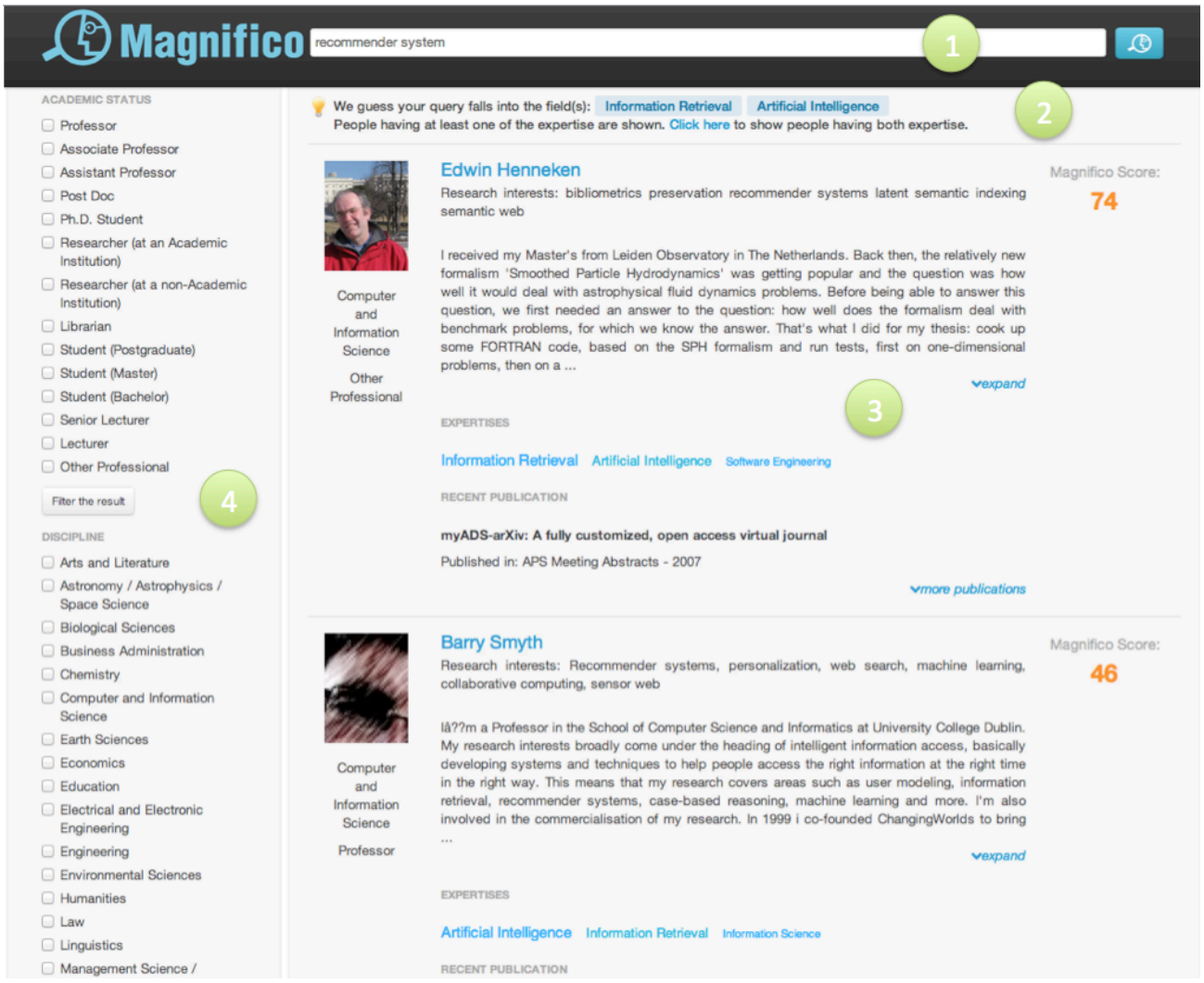

Figure 1: The user interface of the *Magnifico* platform

Part 1 of the user interface is the search field where users enter the queries. After users provide the search query, the inferred research field(s) that the query may be categorized into is shown in part 2 of the page. Keeping users aware of the inferred results can help them to refine the queries if the initial search results are not satisfying. Furthermore, if there are more than one research fields matching the query, people with at least one corresponding expertise are displayed by default. Users are allowed to adjust the search condition and look for people having all the expertise at the same time. For instance, the query "recommender system" is matched with both "Information Retrieval" and "Artificial Intelligence" fields. A set of people with the expertise of either "Information Retrieval" or "Artificial Intelligence" is shown by default. But users can choose to obtain people with both expertises.

The search result is presented in part 3 of the page, sorted by the *Magnifico* score. 10 people are loaded each time to increase the performance and improve the user experience of the system. More people will be loaded if the page is scrolled to the bottom. For each person having a Mendeley profile, the profile photo, main research discipline, academic status, name, research interests, and biographic information are shown, as well as a link to her Mendeley profile page for future contact. As for the authors who have not registered with Mendeley, only the name is presented. In addition to the basic profile information, an expertise cloud is also shown for each person, revealing her top expertise. The research field with bigger font represents stronger expertise. The publications of the person are listed below the expertise cloud. On the right side of part 3, the *Magnifico* score is illustrated, indicating the strength of the expertise in a numerical way.

Finally, *Magnifico* also supports filtering within the search results according to people's academic statuses and research disciplines, as shown in part 4 of the page. A job recruiter might be interested in looking for only master and Ph.D. students with some specific competences. So people with other academic statuses can be removed from the list. At the bottom area of part 4, users are allowed to constrain the search conditions to include or exclude certain disciplines.

CONCLUSIONS AND FUTURE WORK

In this paper, an expert mining platform, *Magnifico*, is presented. A modified TF-IDF approach is used to infer the topics of the search query, and a multi-disciplinary reputation metric is also introduced to compute people's expertise in specific fields. The user interface of the platform is described in the end.

The platform currently only makes use of the metadata in the Mendeley corpus. Integration of the social network data into the existing approach is on our research agenda. It is also planned to adopt other text mining technique such as Latent Dirichlet Allocation model. Also, the evaluation will be conducted to examine the usefulness and usability of the platform.

REFERENCES

1. D. Gillet, E.L.-C. Law, and A. Chatterjee, "Personal learning environments in a global higher engineering education Web 2.0 realm", *Education Engineering*, pp. 897-906, 2010.
2. K. Jack, J. Hammerton, D. Harvey, J. J Hoyt, J. Reichelt, and V. Henning, "Mendeley's reply to the DataTEL challenge", *Procedia Computer Science*, vol. 1, pp. 1-3, 2010.
3. Mendeley API Documentation. http://apidocs.mendeley.com.
4. Chromium Compact Language Detector. http://code.google.com/p/chromium-compact-language-detector.
5. J. Ramos, "Using TF-IDF to determine word relevance in document queries", In *First International Conference on Machine Learning*, New Brunswick:NJ, USA, 2003.
6. D. D. Lewis, "Naïve (Bayes) at forty: the independence assumption in information retrieval", In *10th European Conference on Machine Learning*, pp. 4-15, 1998.
7. Standard Set of English Stop Words. https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords.
8. Bootstrap Framework. http://twitter.github.com/bootstrap.
9. Ruby on Rails Framework. http://rubyonrails.org.