An Expertise Recommender System Based on Data from an Institutional Repository (DiVA)
Milena Angelova, Vishnu Devagiri, Veselka Boeva, Peter Linde, Niklas Lavesson

To cite this version:
Milena Angelova, Vishnu Devagiri, Veselka Boeva, Peter Linde, Niklas Lavesson. An Expertise Recommender System Based on Data from an Institutional Repository (DiVA). Leslie Chan; Pierre Mounier. ELPUB 2018, Jun 2018, Toronto, Canada. <10.4000/proceedings.elpub.2018.17>. <hal-01816680>
An Expertise Recommender System Based on Data from an Institutional Repository (DiVA)

Milena Angelova, Vishnu Manasa Devagiri, Veselka Boeva, Peter Linde and Niklas Lavesson

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

Introduction

Finding experts in academics is an important practical problem, e.g. recruiting reviewers for reviewing conference, journal or project submissions, partner matching for research proposals, finding relevant M. Sc. or Ph. D. supervisors etc. In this work, we discuss an expertise recommender system that is built on data extracted from the Blekinge Institute of Technology (BTH) instance of the institutional repository system DiVA (Digital Scientific Archive). DiVA is a publication and archiving platform for research publications and student essays used by 46 publicly funded universities and authorities in Sweden and the rest of the Nordic countries (www.diva-portal.org). The DiVA classification system is based on the Swedish Higher Education Authority (UKÄ) and the Statistic Sweden's (SCB) three levels classification system. Using the classification terms associated with student M. Sc. and B. Sc. theses published in the DiVA platform, we have developed a prototype system which can be used to identify and recommend subject thesis supervisors in academy.

Related Work

The discovery of expertise is crucial in supporting a number of tasks. For example, finding an appropriate expert when one needs guidance on a subject matter, or needs to fill a vacancy based on relevant expertise, or needs to find research collaborators working
in similar areas etc. The methods used for developing systems that facilitate mining of expertise can be classified into two main categories: methods based on mining unstructured information and methods based on social networking sites.

3 Unstructured information includes emails, Web pages, wiki, reports, etc. Text mining tools are used to index technical terms from unstructured documents, which can be queried to identify subject experts. Some of the important tools using this information include email expertise extraction (e3) system (Krulwich et al. 1996), ContactFinder (Campbell et al. 2003), MIT Expert Finder (Vivacqua 1999), etc. A major limitation of these methods is the authentication of information.

4 Today there are many social networking sites, some specifically for the scientific community, such as Research Gate, Nature Network etc. Experts have to feed in information about their subject expertise, domains, publications, credentials, etc. A major limitation of these methods is in the adding and updating of information.

5 A common drawback to both approaches is that they are based on non-peer-reviewed information provided by the user, i.e. they can be biased.

6 In the recent years research on identifying experts from online data sources, such as the DBLP library, Microsoft Academic Search, Google Scholar Citation, LinkedIn, PubMed etc., has been gradually gaining interest (Abramowicz et al. 2011, Balog and de Rijke 2007, Bozzon et al. 2013, Boeva et al. 2014, Boeva et al. 2016, Hristoskova et al. 2013, Jung et al. 2007, Singh et al. 2013, Stankovic et al. 2011, Tsiorkova and Tourwé 2011, Zhang et al. 2007). An example is a Web-based biomedical expert finding system, proposed in (Singh et al. 2013), which can be applied to identify subject experts and subjects associated with an expert. The system builds and maintains a big repository of biomedical experts by extracting the information about experts’ peer-reviewed articles that are published and indexed in PubMed. Two other interesting approaches using semantic matching and developing trust algorithms are (Charlin et al., 2013) and (Osman et al., 2013). (Charlin et al., 2013) have developed an automated reviewer assignment system that can be used to find suitable reviewers for conference papers. The assessment of reviewers’ expertise is based on their previously published papers and self-assessed expertise about the submissions. (Osman et al., 2013) propose a trust model that calculates the expectation about an agent’s future performance in a given context by assessing both the agent’s willingness and capability through the semantic comparison of the current context in question with the agent’s performance in past similar experiences. A comprehensive survey on the state-of-the-art methods in expert finding and summary of these methods into different categories based on their underlying algorithms and models is presented in (Lin et al. 2017).

7 Institutional repositories have been around for at least 20 years, but it is only recently you find some discussions on how to create new services using text analysis or data mining in institutional repositories or library catalogues (Kerdprasop et al. 2012, Lawrence 2016, Okamoto 2016, Tonon and Fusco 2014) and even more rare are any evidence on actual resources based on these ideas.

8 The use of university repositories is not so common, in spite of their wealth of bibliographic metadata of both local scientific records and student theses records. There are papers describing mining the contents of scientific papers for more effective information discovery and selection (Saggion and Ronzano 2017, Tonon and Fusco 2014, Okamoto 2016). Other areas that are reported in the literature is personalized services for
library user using data mining technology or book recommendation systems for library user based on user profiling (Ding 2017, Jomsri 2014, Ciu et al. 2014). These are the main uses of data mining in university institutional repositories that we can find being reported in the literature so far. One of the reasons for not exploiting institutional repositories on a greater scale could well be that they are often created, administered and run by librarians who generally are not research oriented and especially not towards computer science. Another reason might be that the data in the repositories often is entered by researchers and students themselves and therefore, in many cases, are not quality controlled. An elaborated discussion of this issue can be found in the section “Data”.

**Identifying Experts from Institutional Data Sources**

Many scientists who work on the expertise retrieval problem distinguish two information retrieval tasks: expert finding and expert profiling, where expert finding is the task of finding experts given a topic describing the required expertise (Craswell et al. 2006), and expert profiling is the task of returning a list of topics that a person is knowledgeable about (Balog 2007). In the considered context we need to deal with both expertise retrieval tasks.

**Profiling of Expertise**

An expert profile may be quite complex and can, for example, be associated with information that includes: email address, affiliation, a list of publications, co-authors, but it may also include or be associated with: educational and (or) employment history. This information can be separated into two parts: expert’s personal data and information that describes the expert area of competence. Personal data is used to resolve the problem with ambiguity. This problem refers to the fact that multiple profiles may represent one and the same person and therefore must be merged into a single generalized expert profile. Namely, the process of merging expert profiles is driven by the calculation of the similarity scores between different entities composing the profile, e.g., expert name, affiliation, email address, etc. (Boeva et al., 2012).

The data needed for constructing the expert profiles could be extracted from various Web sources, e.g., LinkedIn, the DBLP library, Microsoft Academic Search, Google Scholar Citation, PubMed etc. In our work, we have used data extracted from DiVA repository to construct the tutor profiles and ontology model. Initially, the extracted data was divided into personal data and a list of subject terms. Personal data has entities, e.g., first and last name of each thesis supervisor, affiliation containing information about university, faculty and department, and also the academic role of the supervisor. Each B. Sc. or M. Sc. thesis presented in DiVA is associated with a list of keywords that are selected and entered by the author in order to describe her/his thesis subject. We have used these keywords and the built ontology to describe the supervisors’ area of competence. The built ontology is a conceptual model of the domain of interest and it is used to attain accurate and topic relevant expert profiles. When a conceptual model is missing then, e.g. the Stanford part-of-speech tagger (Toutanova and Manning, 2000) can be used to annotate the different words in the text collected for each expert (supervisor) with their specific part of speech. The tagger also defines whether a noun is a plural, whether a verb
is conjugated, etc. The annotated text can further be reduced to a set of keywords (tags) by removing all the words tagged as articles, prepositions, verbs, and adverbs.

In view of the above, an expert profile is defined as a list of keywords (domain-specific topics), extracted from the available information, describing her/his expert area. The problem with ambiguity has been resolved using the extracted personal data. The expert profiles with the same names are candidates to be merged into a single expert profile. For instance, if the supervisors’ names and affiliations are identical then their expert profiles are merged into a single one. Most supervisors who are affiliated with BTH have not entered their email addresses in the DiVA repository. Therefore, we have not used the email address as an entity to resolve the problem with ambiguity.

**Expertise Similarity**

An important issue in an expertise retrieval context is to establish a way to quantify how well the area of expertise of an individual expert conforms to a certain subject or to estimate the expertise similarity between two experts. The calculation of expertise similarity is a complicated task, since the supervisor expertise profiles usually consist of domain-specific keywords that describe their area of competence without any information for the best correspondence between the different keywords of two compared profiles. One possibility to measure the expertise similarity between two expert profiles (or between a supervisor profile and a subject profile) is by taking into account the semantic similarities between any pair of keywords that are contained in the two profiles. Several semantic distance and similarity algorithms can be applied within structured terminological resources (Leacock and Chodorow 1998, Rada et al. 1989, Wu and Palmer 1994). In this case, the algorithms count the number of edges (links) between the two keywords (classification terms) in order to compute the relatedness of these terms.

In the current study, we use a modified Wu and Palmer measure, proposed in (Manjula Shenoy et al., 2012), to calculate similarity between two ontological elements. In considered context all expert profiles are described by a list of domain specific terms, where he/she is an expert. Assume that each expert profile \( i \) is described by a list of \( P_i \) keywords. In our experiments, we use the modified Wu and Palmer measure to calculate the semantic similarity between two keywords. Manjula Shenoy et al. propose an improved version of the standard Wu and Palmer measure (Manjula Shenoy et al., 2012). The modified Wu and Palmer semantic similarity measure \( s \) is defined as follows:

\[
s = \frac{(2N_e^{-\lambda L/D})}{N1 + N2},
\]

where \( L \) is the shortest distance between two given concepts, \( D \) is the depth of the ontology, \( N \) is the distance from the least common ancestor to the root, \( N1 \) and \( N2 \) are respectively, the distances from the two considered concepts to the root, and \( \lambda \) is a coefficient that is 0 when the concepts are in the same hierarchy and 1 when they are neighborhood concepts. Namely, the authors consider how far the two considered concepts are semantically and where they are located in the ontology. The modified Wu and Palmer measure (noted by \( s \) herein), finds the shortest path between two concepts in the ontology tree and the depth of the whole ontology. Then the expertise similarity \( S_{ij} \)
between two expert profiles \( i \) and \( j \) \((i \neq j)\), can be defined by the arithmetic mean of semantic similarities between the corresponding keywords, i.e.

\[
S_{ij} = \frac{1}{p_i \cdot p_j} \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} s(k_{il}, k_{jm}),
\]

where \( s(k_{il}, k_{jm}) \) is the semantic similarity between keywords \( k_{il} \) and \( k_{jm} \) calculated accordingly to the definition given above (Manjula Shenoy et al., 2012).

**Expert Finding**

The experts finding task can be viewed as a list completion task, i.e. the user is supposed to provide a small number of example experts who have been used to work on similar problems in the past, and the system has to return experts with similar area of competence. In our context, the user can be a university student, a director of a study programme or other administrative staff and the returned experts are recommended subject thesis supervisors.

Another possibility is to present the domain of interest by several preliminary specified subject categories and then the available experts can be grouped with respect to these categories into a number of disjoint expert areas (clusters) by using some clustering algorithm, as e.g. ones developed in (Boeva et al. 2014, Boeva et al. 2016). In the considered context each cluster of experts can itself be thought as the expertise area of any expert (supervisor) assigned to the cluster. In this case, in order to select the right individuals (thesis supervisors) for a specified task the user may restrict her/his considerations only to those experts who are within the cluster that is identical with (or at least most similar to) the task's subject. The specified subject and the expert area can themselves be described by lists of keywords, i.e. they can be compared by way of similarity measurement.

**Description of the Prototype System and Results**

The developed prototype system has been evaluated and validated on information extracted from the BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted DiVA classification terms are used to build an ontology that conceptualizes the thesis domain supported by the university. The supervisor profiles of the tutors affiliated with the BTH are constructed based on the extracted DiVA data. Each supervisor profile is defined by a list of keywords (classification terms) used in the DiVA theses that have been supervised by the researcher in question to describe her/his area of expertise.

**Data**

The data set consists of 2216 records of student theses published between 2010 and 2017. One issue with data entered during a long interval like this is that the quality of the records varies. In the case of our sample data this certainly is true. In 2015 the BTH repository joined the national Swedish consortia DiVA-system. At around the same time
the library started supplying records to a national portal “SwePub” intended to become a national tool and a resource for the collection of publication data for bibliometric analyzes and data processing (SwePub). The DiVA migration and having SwePub harvesting our records meant that the library needed to focus much more attention on the quality of the records. The amount of meta data attached to every record increased.

Records entered by students or staff themselves usually have quality issues. In our case we have quality issues with records entered before the second half of 2015. Beginning in the fall of 2015 records are always quality checked by staff from the library or the university departments.

Another data issue is knowing what expert in the used data set still is affiliated with BTH. In a data set covering eight years many supervisors have come and gone. In our case we needed a big data set to create and validate a prototype system so the issues discussed above are not really a problem. But if you want to put an Expertise Recommender System into everyday use issues like these must be paid close attention to.

Ontology Description

We have defined our ontology as a formalization of concepts and relations between them. The domain concepts are described by classes and the relations between them is defined as “isSubclassOf”. Main focus in our ontology are classes and their sub-classes. The model in this study is created using National Subject Categories in DiVA (Linköping University Library, 2011) that is standard for the Swedish listing of research subjects 2011 and classification terms from extracted data. National Subject Categories in DiVA are separated into three levels. Namely, the main categories (the highest level) are Agricultural Sciences, Engineering and Technology, Humanities, Medical and Health Sciences, Natural Sciences, and Social Science. These main categories are located at the first level of the ontology. The others two levels are described by the sub-subject areas that belong to the each of main categories. The second level has 26 subject categories and the third level consists of 257 categories, respectively. From the extracted data we have taken the subject terms (specific domain keywords) and added them to the structure of ontology model as classes. In that way, the created ontology tree has a four level depth.

Metrics

Silhouette Index (Rousseeuw, 1987) is a cluster validity index that is used to judge the quality of any clustering solution \( C = \{C_1, C_2, \ldots, C_k\} \). Suppose that the considered data set contains the attribute vectors of \( m \) objects. Then the SI is defined as

\[
s(C) = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{b_i - a_i}{\max\{a_i, b_i\}} \right),
\]

where \( a_i \) represents the average distance of object \( i \) to the other objects of the cluster to which the object is assigned, and \( b_i \) represents the minimum of the average distances of object \( i \) to object of the other clusters. The values of Silhouette Index vary from -1 to 1.
Results

As it was mentioned above the test data is downloaded from BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted data has been used to construct the supervisor profiles and ontology model. Each supervisor profile is defined by a list of keywords (domain-specific terms). After resolving the problem with ambiguity, the set of supervisors has been reduced to 375 expert profiles.

We have studied two different experiment scenarios. In the first scenario, the user is supposed to provide an example supervisor and the developed recommender system will return a list of experts (supervisors) with close (similar) expertise. In this scenario, initially the system calculates the expertise similarity scores between the given supervisor and the all other supervisors presented in the system by using the expertise similarity formula described in section “Expertise Similarity”. Then the system returns a ranked list of similar experts who are recommended subject thesis supervisors. For instance, we have tested the system by using an example supervisor whose expertise can be described by the following subject terms: database, performance, usability, web server, cloud computing and Amazon web services. The system has returned a list of the supervisors ranked with respect to the similarity of their expert profiles to the profile of the example supervisor. In Table 1, we list the ten top ranked supervisors who have been recommended by the system. The calculated expertise similarity scores between the example supervisor and the all other supervisors presented in the system are in the range between 0.046 and 0.77. The expertise of each top recommended supervisor is described by the keywords found in her/his expert profile (the second column in Table 1). As one can notice the all top ranked supervisors have expertise that is overlapped to different extend with the competence of the given example supervisor. For example, the expertise of supervisor 1 covers usability subject from the example expert profile while supervisor 8 has competence overlap in two subjects: database and performance.

Table 1 The 10 top ranked subject thesis supervisors recommended by the system in the first experiment scenario.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Keywords</th>
<th>Expertise similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>user experience, usability</td>
<td>0.770</td>
</tr>
<tr>
<td>2</td>
<td>privacy, security, cloud computing</td>
<td>0.763</td>
</tr>
<tr>
<td>3</td>
<td>cloud computing, security metrics, security threats, security measurement frameworks</td>
<td>0.760</td>
</tr>
<tr>
<td>4</td>
<td>procedural city generation, perlin noise, performance, game content</td>
<td>0.760</td>
</tr>
<tr>
<td>5</td>
<td>machine learning, parallel computing, multiprocessor, performance</td>
<td>0.760</td>
</tr>
<tr>
<td>6</td>
<td>mobile, power, consumption, android, native, web, enterprise service bus, performance, framework</td>
<td>0.754</td>
</tr>
</tbody>
</table>
In the second scenario, the set of supervisors is clustered into groups of experts with similar expertise by applying k-means partitioning algorithm. The optimal number of clusters is determined by clustering the set of supervisors applying k-means for different k and evaluating the obtained solutions by the Silhouette Index (SI). We have applied k-means for k between 2 and 257 (the maximum domain-specific classes in ontology). The selected optimal value for k is 16 and the corresponding average SI score given on the clustering solution generated for k = 16 is 0.38. In this scenario, in order to select the right individuals for a specified thesis subject the user may restrict her/his considerations only to those supervisors who are within the cluster that is identical with (or at least most similar to) the given subject. The experts in the selected cluster can be ranked with respect to the similarity of their expert profiles to the specified subject. For example, if we need a supervisor with expertise in database and parallel computing we will use cluster 0 (see Table 2), in which all experts have competence either in one or in both fields.

Table 2 Grouping of the 10 top ranked supervisors recommended by the system in the first experiment scenario (ones listed in Table 1).
We have checked how the 10 top ranked supervisors returned by the system in the first experiment scenario has been distributed between the generated clusters. The obtained grouping is presented in Table 2. As we can see the supervisors who have been grouped together have a similar area of competence. For instance, supervisors 3, 4, 7 and 8 are distributed in one and the same cluster (cluster 0). The main expertise of supervisors 3 and 4 is in cloud computing while supervisors 7 and 8 both have competence in database. The supervisors 2, 5, 6 and 10 are grouped in the cluster 3. For example, the supervisors 5 and 6 have expertise in performance and the others are in cloud computing and video streaming. However, supervisors 1 and 9 are in a different cluster (cluster 15 in Table 2), because as we can see in Table 1 they have very unique expertise in comparison with the others.

It is interesting to notice that the generated clustering solution partitions the supervisors into 16 disjoint clusters (subject categories), i.e. each supervisor belongs to only one cluster. However, as it can be seen in Table 2 there are subject terms that participate in the description of more than one subject category, e.g., parallel computing is used for clusters 0 and 3 (the third column in Table 2), i.e. the obtained subject categories can have an overlap on some topics.

Conclusion and Future Work

The main contributions of this work are:

i. a prototype system for ranking potential thesis supervisor candidates based on their previous supervision experience and expertise retrieval approaches; and

ii. an evaluation of the prototype with real-world data extracted from the BTH instance of the institutional repository system DiVA. The proposed system is based on online information, can be updated regularly, and uses standardized subject vocabulary, i.e., DiVA classification terms associated with each thesis. It can be implemented by any university that uses an institutional repository with a controlled vocabulary including keyword fields in the records, i.e. it is general.

For future work, we aim to pursue further evaluation and validation of the developed expert recommended system on bigger datasets and also on data extracted from other DiVA university repositories. Our future plans also involve using additional information about the supervisors’ expertise, e.g., extracting information about the supervisors’ research publications, in order to build richer expert profiles and ontology model.
BIBLIOGRAPHY

References


SwePub för analys och bibliometri. (webportal). http://bibliometri.swepub.kb.se/bibliometri


ABSTRACTS

Finding experts in academics is an important practical problem, e.g. recruiting reviewers for reviewing conference, journal or project submissions, partner matching for research proposals, finding relevant M. Sc. or Ph. D. supervisors etc. In this work, we discuss an expertise recommender system that is built on data extracted from the Blekinge Institute of Technology (BTH) instance of the institutional repository system DiVA. The developed prototype system is evaluated and validated on information extracted from the BTH DiVA installation, concerning thesis supervision of researchers affiliated with BTH. The extracted DiVA classification terms are used to build an ontology that conceptualizes the thesis domain supported by the university. The supervisor profiles of the tutors affiliated with the BTH are constructed based on the extracted DiVA data. These profiles can further be used to identify and recommend relevant subject thesis supervisors.

INDEX

Keywords: data mining, DiVA, expertise retrieval, knowledge management, natural language processing

AUTHORS

MILENA ANGELOVA
Technical University of Sofia-branch Plovdiv, Bulgaria
mangelova@tu-plovdiv.bg

VISHNU MANASA DEVAGIRI
Blekinge Institute of Technology, Sweden
vishnu.manasa26@gmail.com

VESELKA BOEVA
Blekinge Institute of Technology, Sweden
veselka.boeva@bth.se
(corresponding author)

PETER LINDE
Blekinge Institute of Technology, Sweden
peter.linde@bth.se

NIKLAS LAVESSON
Blekinge Institute of Technology, Sweden
niklas.lavesson@bth.se