Advanced Data-driven Techniques for Mining Expertise

Milena Angelova[1], Veselka Boeva[2] and Elena Tsiporkova[3]
[1]Technical University of Sofia-branch Plovdiv, Bulgaria mangelova@tu-plovdiv.bg
[2]Blekinge Institute of Technology, Karlskrona, Sweden
[3]Sirris, The Collective Center for the Belgian technological industry, Brussels, Belgium

Abstract
In this work, we discuss enhanced techniques that optimize expert representation and identify subject experts via clustering analysis of the available online information. We use a weighting method to assess the levels of expertise of an expert to the domain-specific topics. In this context, we define a way to estimate the expertise similarity between experts. Then the experts finding task is viewed as a list completion task and techniques that return similar experts to ones provided by the user are considered. In addition, we discuss a formal concept analysis approach for clustering of a group of experts with respect to given subject areas. The produced grouping of experts can further be used to identify individuals with the required competence.

Keywords
Data mining, expert finding, health science, knowledge management.

1 INTRODUCTION

Nowadays, organizations search for new employees not only relying on their internal information sources, but they also use data available on the Internet to locate required experts. As the data available is very dispersed and of distributed nature, a need appears to support this process using IT-based solutions, *e.g*., information extraction and retrieval systems, especially expert finding systems. Expert finding systems however, need a lot of information support. On one hand, the specification of required "expertise need" is replete with qualitative and quantitative parameters. On the other hand, the expert finders need to know whether a person who meets the specified criteria exists, how extensive his/her knowledge or experience is, whether there are other persons who have the similar competence, how he/she compares with others in the field, *etc*. Consequently, techniques that gathers and makes such information accessible are needed.

Many practical scenarios of organizational situations that lead to expert seeking have been extensively presented in the literature. For instance, Autonomy [1] analyses users' search and publication histories to determine concepts that are indicative of their expertise. Yenta [2] and Tacit KnowledgeMail [3] determine user expertise from email message traffic. Referral Web from AT&T [4] provides access to experts across an expertise or community, aiming to make the basis for referral transparent to the user. In recruitment industry, the problem of finding expertise is an one of seeking for job candidates given the required skills as well as some additional information, such as, location and/or company names [5]. Several Web-based expert seeking tools that support both type players at the job market have recently appeared [6][7]. For instance, in [6], a personalized job seeking for an applicant is proposed by given him/her benchmark against the current job market.

Expert finders are usually integrated into organizational information systems, such as knowledge management systems, recommender systems, and computer supported collaborative tasks. Initial approaches propose tools that rely on people to self-assess their skills against a predefined set of keywords, and often employ heuristics generated manually based on current working practice [8]. Later approached try to find expertise in specific types of documents, such as e-mails [9][10] or source code [11]. Instead of focusing only on specific document type systems that index and mine published intranet documents as sources of expertise evidence are discussed in [12]. In the recent years, research on identifying experts from online data sources has been gradually gaining interest [13][14][15][16][17].

In this work, we discuss enhanced techniques that optimize expert representation and identify subject experts via clustering analysis of the available online information. In [23], we have proposed a weighting method to assess the levels of expertise of an expert to the domain-specific topics. In this context, we have further defined a way to estimate the expertise similarity between experts. Then the experts finding task is viewed as a list completion task and techniques that return similar experts to ones provided by the user are considered. In addition, we have proposed a formal concept analysis approach for clustering of a group of experts with respect to given subject areas [33]. The produced grouping of experts can further be used to identify individuals with the required competence. The proposed expert finding techniques have been evaluated on data extracted from PubMed repository.

.

2 PROPOSED SOLUTIONS

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 [18], and *expert profiling* is the task of returning a list of topics that a person is knowledgeable about [19].

In this work, we consider data-driven techniques that deal with both expertise retrieval tasks. Initially, we need to describe the expertise of each involved person by creating his/her expert profile, *i.e.* each person is associated by a list of subjects he/she is an expert in.

2.1 Expert profiling

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. There exist several open tools for extracting data from public online sources. For instance, Python LinkedIn is a tool which can be used in order to execute the data extraction from LinkedIn. In addition, the Stanford part-of-speech tagger [20] can be used to annotate the different words in the text collected for each expert with their specific part of speech. Next to recognizing the part of speech, the tagger also defines whether a noun is plural, whether a verb is conjugated, etc. Further the annotated text can be reduced to a set of keywords (tags) by removing all the words tagged as articles, prepositions, verbs, and adverbs. Practically, only the nouns and the adjectives are retained.

However, an expert profile may be quite complex and can, for example, be associated with information that includes: e-mail address, affiliation, a list of publications, co-authors, but it may also include or be associated with: educational and (or) employment history, the list of LinkedIn contacts etc. All this information could be separated into two parts: the expert's personal data and information that describes the competence area of expert.

The expert's personal data can be 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, *e.g.*, the clustering algorithm discussed in [21] can be applied for this purpose. A different approach to the ambiguity problem has been proposed in [22]. Namely, the similarity between the personal data (profiles) of experts is used to resolve the problem with ambiguity. The split and merge of expert profiles is driven by the calculation of similarity measure between the different entities composing the profile, *e.g.* expert name, email address, affiliations, co-authors names etc. Thus the similarity between the personal data of two expert profiles is defined by the weighted mean of similarities between the corresponding fields of their profiles [22].

In [23], we use a Dynamic Time Warping (DTW) based approach to deal with the ambiguity issue. In general, the DTW alignment algorithm finds an optimal match between two given sequences (*e.g.*, time series) by warping the time axis iteratively until an optimal matching (according to a suitable metric) between the two sequences is found [24]. Due to its flexibility, DTW is widely used in many scientific disciplines and business applications as *e.g.*, speech processing, bioinformatics, matching of one-dimensional signals in the online hand writing communities etc. A detail explanation of DTW algorithm can be found in [24][25].

In view of the above, an expert profile can be defined as a list of keywords (domain-specific topics), extracted from the available information about the expert in question, describing her/his subjects of expertise. Assume that n different expert profiles are created in total and each expert profile i (i = 1, 2, ..., n) is represented by a list of p_i keywords.

2.2 Assessing of expertise

An expert may have more extensive knowledge or experience in some topics that in others and this should be taken into account in the constructions of expert profiles. Thus the gathered information about each individual expert can further be analyzed and used to access her/his levels of expertise to the different topics that compose her/his expert profile.

There is no standard or absolute definition for accessing expertise. This usually depends not only on the application area but also on the subject field. For instance, in the peer-review setting, appropriate experts (reviewers, committee members, editors) are discovered by computing their profiles, usually based on the overall collection of their publications [26]. However, the publication quantity alone is insufficient to get an overall assessment of expertise. To incorporate the publication quality in the expertise profile, Cameron used the impact factor of publications' journals [26]. However, the impact factor in itself is arguable [27][28]. Therefore, Hirsch proposed another metric, the "HIndex", to rank individuals [29]. However, this index works fine only for comparing scientists working in the same field, because citation conventions differ widely among different fields [29]. Afzal et al. proposed an automated technique which incorporates multiple facets in providing a more representative assessment of expertise [30]. The developed system mines multiple facets for an electronic journal and then calculates expertise' weights.

In [23], we have proposed a weighting method to assess the levels of expertise of an expert to the domain-specific topics. Namely, weights are used to access the relative levels of knowledge or experience an individual has in the topics he/she has shown to have an expertise. Let us suppose that a weighting method appropriate to the respective area is used and as a result each keyword (domain-specific topic) k_{ij} of expert profile i (i = 1, …, n) is associated with a weight w_{ij}, expressing the relative level (intensity) of expertise the expert in question has in the topic k_{ij}, *i.e.* $\sum_{j=1}^{p_i} w_{ij} = 1$ and $w_{ij} \in (0, 1]$ for i = 1, …, n.

In this way, each expert can be presented by a richer expert profile describing the topics (keywords) in which he/she is an expert plus the levels (weights) of knowledge or

experience he/she has in the different topics. Namely, each expert is represented by two components: a list of keywords (topics) and a vector of weights expressing the relative levels of expertise the expert has in the different topics.

2.3 Expertise similarity

The calculation of expertise similarity is a complicated task, since the expert 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. Therefore, it is proposed in [22] to measure the similarity between two expertise profiles as the strength of the relations between the semantic concepts associated with the keywords of the two compared profiles. Another possibility to measure the expertise similarity between two expert profiles is by taking into account the semantic similarities between any pair of keywords that are contained in the two profiles.

Accurate measurement of semantic similarity between words is essential for various tasks such as, document (or expert) clustering, information retrieval, and synonym extraction. Semantically related words of a particular word are listed in manually created general-purpose lexical ontologies such as WordNet. WordNet is a large lexical database of English [31][32]. Initially, the WordNet networks for the four different parts of speech were not linked to one another and the noun network was the first to be richly developed. This imposes some constraints on the use of WordNet ontology. Namely, most of the researchers who use it limit themselves to the noun network. However, not all keywords representing the expert profiles are nouns. In addition, the algorithms that can measure similarity between adjectives do not yield results for nouns hence the need for combined measure. Therefore, a normalized measure combined from a set of different similarity measures is defined and used in [33] to calculate the semantic relatedness between any two keywords. In the considered context the expertise similarity task is additionally complicated by the fact that the competence of each expert is represented by two components: a list of keywords describing her/his expertise and a vector of weights expressing the relative levels of knowledge/expertise the expert has in the different topics.

Let s be a similarity measure that is suitable to estimate the semantic relatedness between any two keywords used to describe the expert profiles in the considered domain. Then the expertise similarity S_{ij} between two expert profiles i and j ($i \neq j$), can be defined by using the weighted mean of semantic similarities between the corresponding keywords

$$S_{ij} = \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} \cdot s(k_{il}, k_{jm}), \qquad (1)$$

where $W_{lm} = w_{il}.w_{jm}$ is a weight associated with the semantic similarity $s(k_{il}, k_{jm})$ between keywords k_{il} and k_{jm}, and $W_{lm} \in (0, 1]$ for $i = 1, \ldots, p_i$ and $m = 1, \ldots, p_j$. It can easily be shown that $\sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} = 1$.

2.4 Identifying experts through clustering

In [33], we have proposed a formal concept analysis approach for grouping a given set of experts with respect to pre-defined subject areas. Initially, the domain of interest is described at some level of abstraction by partitioning the domain to a number of subject areas. Next each expert is represented by a vector of contributions (membership degrees) of the expert to the different areas. This defines an overlapping partition, which is further analysed and refined into a disjoint one by applying Formal Concept Analysis (FCA).

A conceptual model of the domain of interest, such as a thesaurus, a taxonomy etc., can be available. In this case, usually a set of subject terms (topics) arranged in hierarchical manner (tree structures) is used to represent concepts in the considered domain. Further it can be supposed that the tree structure describing the considered domain has k main branches (broad subject categories). Another possibility to represent the domain of interest at a higher level of abstraction is to partition the set of all different keywords used to define the expert profiles into k groups (main subject areas). The latter idea has been proposed and applied in [34]. Initially, a common set of all different keywords is formed by pooling the keywords of all the expert profiles. Then the semantic distance between each pair of keywords is calculated and the keywords are partitioned by applying a selected clustering algorithm.

As discussed above, the domain of interest can be presented by k main subject categories $C_1, C_2, \ldots, C_k$. Let us denote by b_{ij} the number of keywords from the expert profile of expert i that belong to category (subject area) C_j. In [33], we have assumed that each expert i ($i = 1, 2, \ldots, n$) is described by only a list of the domain-specific topics (keywords) in which he/she is an expert. Then this representation can be converted into a vector $e_i = (e_{i1}, e_{i2}, \ldots, e_{ik})$, where $e_{ij} = b_{ij}/p_i$ ($j = 1, 2, \ldots, k$) and p_i is the total number of keywords in the expert profile.

In this way, each expert i is represented by a k-length vector of membership degrees of the expert to k different subject categories, *i.e.* the above procedure generates a *fuzzy* clustering. Thus an expert will have a membership degree of 1 to a certain subject area only in case all the keywords of her/his expert profile are assigned to the category in question. The resulting fuzzy partition can easily be turned into a *crisp* one by assigning to each pair (expert, area) a binary value (0 or 1), *i.e.* for each subject area we can associate those experts who have membership degrees greater than a preliminary given threshold (*e.g.* 0.5). Notice, this partition is not guaranteed to be disjoint in terms of the different subject area, since there will be experts who will belong to more than one subject category.

The above overlapping partition is further analyzed and refined into a disjoint one by applying Formal Concept Analysis. FCA is a principled way of automatically deriving a hierarchical conceptual structure from a collection of objects and their properties [35]. The approach takes as input a matrix (referred to formal context) specifying a set of objects and the properties thereof, called attributes.

In our case, a (formal) **context** consists of the set of the n experts, the set of main categories $\{C_1, C_2, \ldots, C_k\}$ and an indication of which experts are associated with which subject category. Thus the context is described as a matrix, with the experts corresponding to the rows and the categories corresponding to the columns of the matrix, and a value 1 in cell (i, j) whenever expert i is associated with (has expertise in) subject area C_j. Subsequently, a (formal) **concept** for this context is defined to be a pair (X, Y) such that

- X is a subset of experts and Y is a subsets of subject areas, and every expert in X belongs to every area in Y
- for every expert that is not in X, there is a subject area in Y that does not contain that expert
- for every subject area that is not in Y, there is an expert in X who is not associated with that area.

The family of these concepts obeys the mathematical axioms defining a **concept lattice** [35]. The built lattice consists of concepts where each one represents a subset of experts belonging to a number of subject areas. The set of all concepts partitions the experts into a set of disjoint expert areas.

Evidently, the produced grouping of experts facilitate the identification of individuals with the required competence. For instance, if we need to recruit experts who have expertise simultaneously in two subject categories, we can directly locate those who belong to the concept that unites the corresponding categories. In addition, such a grouping of experts can be performed with respect to any set of subject areas describing the domain of interest, *e.g.*, the experts could be clustered on a lower level of abstraction by using more specific topics. It is even possible to further produce a grouping of experts belonging to a particular concept around topics specifying the subject areas associated with this concept.

2.5 Finding similar experts

The experts finding task can also 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 similar experts. In [23], we have proposed techniques that return similar experts to ones provided by the user.

The concept of expertise spheres has been introduced in [22]. Conceptually, these expertise spheres are interpreted as groups of experts who have strongly overlapping competences. In other words, the expertise sphere can be considered as a combination of pieces of knowledge, skills, proficiency etc. that collectively describe a group of experts with similar area of competence. Consequently, the user may find experts with the required expertise by entering the name(s) of example expert(s) and the system will return a list of experts with close (similar) expertise by constructing the expertise sphere of the given expert(s).

In order to build an expertise sphere of an expert it is necessary to identify experts with similar area of competence, *i.e.* for each example expert i a list of expert profiles which exhibit at least minimum (preliminary defined) expertise similarity with her/his expert profile needs to be generated. An expert profile j will be included in the expertise sphere of i if the following inequality holds $S_{ij} \geq T$, where $T \in (0, 1)$ is a preliminary defined threshold. The experts identified can be ranked with respect to their expertise similarities to the example expert.

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.* [33][34]. In the considered context each cluster of experts can itself be interpreted as an expertise sphere. Namely, it can be thought as the expertise area of any expert assigned to the cluster and evidently, the all assigned experts are included in this sphere. In this case, in order to select the right individuals for a specified task the user may restrict her/his considerations only to those experts who are within the expert area (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 (subject profiles), *i.e.* they can be compared by way of similarity measurement. In this scenario, weights can also be introduced by allowing the user to express her/his preferences about the relative levels of expertise the experts in query should have in the specified topics. In addition, the subject profiles that are used to present the different clusters of experts can also be supplied with weights. The experts in the selected cluster can be ranked with respect to the similarity of their expert profiles to the specified subject profile.

In case of a newly extracted (registered, discovered) expert we can classify him/her into one of the existing clusters of experts by determining his/her expertise sphere. Namely we initially calculate the expert's expertise spheres with respect to any of the considered expert areas. Then the expert in question is assigned to that cluster of experts for which the corresponding expertise sphere has the largest cardinality, *i.e.* the overlap between the two sets of experts is the highest.

3 EVALUATION AND DISCUSSION

3.1 PubMed data

The data needed for constructing the expert profiles are extracted from PubMed, which is one of the largest repositories of peer-reviewed biomedical articles published worldwide. Medical Subject Headings (MeSH) is a controlled vocabulary developed by the US National

Library of Medicine for indexing research publications, articles and books. Using the MeSH terms associated with peer-reviewed articles published by Bulgarian authors and indexed in the PubMed, we extract all such authors and construct their expert profiles. An expert profile is defined by a list of MeSH terms used in the PubMed articles of the author in question to describe her/his expertise areas.

3.2 Metrics

Unfortunately, large data collections such as *e.g.* LinkedIn, the DBLP library, PubMed etc. contain a substantial proportion of noisy data and the achieved degree of accuracy cannot be estimated in a reliable way. Accuracy is most commonly measured by precision and recall. Precision is the ratio of true positives, *i.e.* true experts in the total number of true experts in a given domain. However, determining the total number of true experts in a given domain is not feasible.

In the current work, we use *resemblance r* and *containment c* to compare the expertise retrieval solutions generated on a given set of experts by using the expertise retrieval techniques discussed in Section 2.5.

Let us consider two expertise retrieval solutions $S = \{S_1, S_2, \dots, S_k\}$ and $S' = \{S'_1, S'_2, \dots, S'_k\}$ of the same set of experts. Then the similarity between two expertise retrieval results S'_i and S_i, which are constructed for the same example expert, can be assessed by *resemblance r*:

$$r(S'_i, S_i) = |S'_i \cap S_i| / |S'_i \cup S_i|, \quad (2)$$

where S_i and S'_i, i = 1, 2,…, k, are corresponding expertise retrieval results. The first solution S is generated on the considered data set without taking into account the expert levels of expertise in different topics while the second one S' is a solution built by applying the proposed weighting method.

We also use *containment c* that assesses how S'_i is a subset of S_i:

$$c(S'_i) = |S'_i \cap S_i| / |S'_i| \quad (3)$$

The values of *r* and *c* are in the interval [0, 1].

We also use *Silhoutte Index* (SI) to evaluate the quality of the cluster solution generated by the FCA-based approach considered in Section 2.4. Silhouette Index is a cluster validity index that is used to judge the quality of any clustering solution $C = \{C_1, C_2, \dots, C_k\}$ of the considered data set, which contains the attribute vectors of m objects. Then the SI is defined as

$$s(C) = 1 \Big/ m \sum_{i=1}^{m} (b_i - a_i) \Big/ \max\{a_i, b_i\}, \quad (4)$$

where a_i represents the average distance of objects *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 value of SI from -1 to 1 and higher value indicates better clustering results.

3.3 Implementation and availability

Publications originating from Bulgaria have been downloaded in XML format from the Entrez Programming Utilities (E-utilities). The E-utilities are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene Nuccore and Protein. The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including biomedical literature. To access these data, a piece of software first makes an API call to E-Utilities server, then retrieves the results of this posting, after which it processes the data as required. Thus the software can use any computer language that can send a URL to the E-utilities server and interpret the XML response.

For calculation of semantic similarities between MeSH headings, we use MeSHSim which is an R package. It also supports querying the hierarchy information of a MeSH heading and information of a given document including title, abstraction and MeSH headings [36].

In our experiments, we have applied the DTW-based algorithm to resolve the problem with ambiguity. For this purpose, we have used a Python library cdtw. It proposes a DTW algorithm for spoken word recognition which is experimentally shown to be superior over other algorithms [37].

3.4 Results and discussion

We have extracted a set of 4343 Bulgarian authors from the PubMed repository. After resolving the problem with ambiguity the set is reduced to one containing only 3753 different researchers. Then each author is represented by two components: a list of all different MeSH headings used to describe the major topics of her/his PubMed articles and a vector of weights expressing the relative levels of expertise the author has in the different MeSH terms composing her/his profile. The weight of a MeSH term that is presented in a particular author profile is the ratio of repetitions, *i.e.* the repetitions of the MeSH term in the total number of MeSH terms collected for the author. This weighting technique could additionally be refined by considering the MeSH terms annotating the recent publications of the authors as more important (*i.e.* assigning higher weights) than those met in the old ones. This idea is not implemented in the current experiments.

Experts	MeSH headings
1	Kidney Transplantation; Liver Transplantation
2	Health Behavior
3	Drinking; Health Behavior; Health Knowledge, Attitudes, Practice; Program Evaluation
4	Models, Biological; Temperature; Models, Neurological; Water
5	Computer Simulation; Models, Molecular; Protons; Thermodynamics; Molecular Conformation
6	Vibration; Models, Molecular; Infrared Rays; Hydrogen Bonding
7	Monte Carlo Method; Models, Theoretical; Phase Transition; Thermodynamics
8	Photosynthesis; Quantum Theory
9	Health Behavior; Decision Support Techniques; . . .(more than 20 MeSH terms)
10	Polymorphism, Genetic

Table 1 Expert MeSH heading profiles.

Experts	MeSH heading weights
1	0.5; 0.5
2	1
3	0.25; 0.25; 0.25; 0.25
4	0.166; 0.333; 0.166; 0.333
5	0.285; 0.285; 0.142; 0.142; 0.142
6	0.5; 0.166; 0.166; 0.166
7	0.428; 0.285; 0.142; 0.142
8	0.75; 0.25
9	0.022;...; 0.045; . . . ; 0.068; . . . ; 0.25
10	1

Table 2 Expert MeSH heading weights.

Examples of 10 expert MeSH heading profiles can be seen in Table 1. The corresponding weight vectors calculated as it was explained above can be found in Table 2.

We build expertise spheres of the ten example experts whose profiles are given in Table 1. Initially, we construct the expertise spheres of these authors by applying the weighting method discussed in Section 2.2. Respectively, the expertise spheres of the same authors without taking into account the intensity of their expertise in the different MeSH topics containing in their profiles are also produced. Next the resemblance *r* and the containment *c* are used to compare the two expertise retrieval solutions generated on the set of extracted Bulgarian PubMed authors for the example expert profiles.

Figure 1 depicts *r* and *c* scores which have been calculated on the expertise retrieval results produced for the example experts by identifying for each expert profile a fixed number (50) of expert profiles that are most similar to the given one. As one can notice the obtained results are quite logical. Namely, the returned expertise retrieval results are identical ($r = 1$ and $c = 1$) when the experts have equally distributed expertise in the different MeSH headings presented in their profiles (*e.g.*, see experts: 1, 2, 3 and 10). However, in the other cases (see experts: 4, 5, 6, 7 and 8) the resemblance between the corresponding expertise retrieval results is not very high (maximum 0.4). Evidently, the produced expertise retrieval results can be significantly changed by using a weighting method for assessing expert expertise. The latter is also supported by the results generated for the containment *c*.

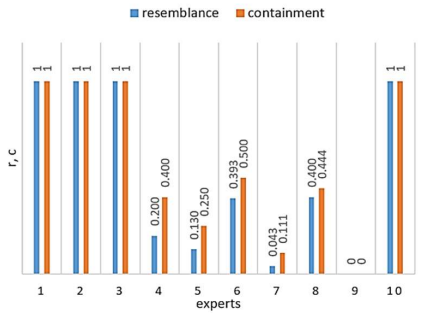

Figure 1 *r* and *c* scores calculated on the expertise retrieval results that are generated for the example experts given in Table 1 by selecting for each expert profile a fixed number of the most similar expert profiles.

The MeSH headings are grouped into 16 broad categories. We have produced a grouping of all the extracted authors with respect to these subject categories by applying the formal concept analysis clustering approach explained in Section 2.4. In this experiment, we have assumed that each author is described by only a list of MeSH terms. Next each author is further represented by a 16-length vector of membership degrees (contributions) of the expert to the different categories. The membership degree to the category is calculated as a ratio between the number of MeSH headings from the author profile that belong to the category and the total number of headings in her/his profile. The calculated membership degrees are turned into binary values by using a preliminary determined threshold. In the considered context the threshold is set to be equal to the median of all different membership degrees. Thus for each subject category we have associated those authors who have membership degrees greater than the determined threshold. Ten of the authors have not been assigned to any category, since there are no membership degrees above the calculated threshold in their profiles. Then a formal context presented by a 3753×16 matrix, with the authors corresponding to the rows and the subject categories corresponding to the columns is built. Finally, a formal concept lattice for the built context is generated by using a data mining prototype Lattice Miner. It produces a lattice of 234 concepts. The non-empty concepts are 198, where each one represents a

subset of authors who belong to a number of subject categories. Thus the extracted Bulgarian health science experts are partitioned into 198 disjoint expert areas with respect to the main MeSH categories. 2166 researchers have been partitioned among 14 singleton concepts, 10 authors belong to the empty concept and the other 1587 researchers demonstrate multiple expertise. The number of authors partitioned into the main MeSH categories (singleton concepts) are given in Table 3.

Category label	Category name	Number of authors
A	Anatomy	45
B	Organisms	101
C	Diseases	68
D	Chemicals and Drugs	158
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment	663
F	Psychiatry and Psychology	97
G	Phenomena and Processes	797
H	Disciplines and Occupations	38
I	Antropology, Education, Socialogy and Social Phenomena	14
J	Tehnology, Industry, Arguculture	20
K	Humanities	2
L	Information Science	37
M	Named Groups	1
N	Health Care	125

Table 3 Number of authors partitioned into the main MeSH categories (singleton concepts).

Evidently, the produced grouping of experts well capture the expertise distribution in the considered domain with respect to the main subjects. In addition, it facilitates the identification of individuals with the required competence. For instance, if we need to recruit researchers who have expertise simultaneously in 'Phenomena and Processes' and 'Health care' categories, we can directly locate those who belong to the concept that unites the corresponding categories ({G, N}). Selected non-singleton concepts are given in Table 4. Most of these concepts unite two categories, *i.e.* the corresponding authors are active in two scientific areas. Logically the number of authors who have expertise in more than two subject categories is not very high.

It is difficult to evaluate the obtained expert partitioning as there are no benchmark ones available. Therefore, we have conducted another experiment in [33]. It performs the semantic-aware expert clustering algorithm, proposed in [34], with our test data. Initially, the constructed expert profiles represented by the 16-length vectors of membership degrees are used to calculate the Euclidean distance between each pair of vectors. Then the authors are clustered by using *k*-means clustering algorithm. However, in order to determine the optimal number of clusters for the considered set of experts we have initially applied *k*-means clustering algorithm for different values of *k* and then we have evaluated the obtained clustering solutions by SI. In comparison to the partitioning algorithms as *k*-means the FCA-based approach does not need prior knowledge about the optimal number of clusters in order to produce a good clustering solution. Notice that the SI score generated on the expert clustering produced by the proposed approach is 0.698.

United categories	Number of authors
{G, N}	106
{E, N}	55
{C, G}	59
{E, L}	36
{F, N}	23
{F, I}	12
{E, G, N}	56
{E, H, J, L}	8
{G, H, L, N}	6
{E, G, I, L, N}	11
{F, G, H, I, N}	7

Table 4 Number of authors partitioned into united MeSH categories (non-singleton concepts)

4 SUMMARY

In this paper, we have discussed enhanced data-driven techniques for expert representation and identification. The proposed techniques have been tested and evaluated on data extracted from PubMed repository.

For future work, we aim to pursue further evaluation, validation and refinement of the discussed expert mining techniques on richer data coming from different application areas, subject fields and online sources, *e.g.* such as LinkedIn, Google Scholar, the DBLP library, Microsoft Academic Search, etc.

5 REFERENCES

[1] Autonomy Technology White Paper (http://www.autonomy.com)

[2] Foner, L. “Yenta: a multi-agent referral system for matchmaking system”, Proceedings of the First International Conference on Autonomous Agents, Marina Del Ray, CA, 1997.

[3] Tacit Knowledge Systems’ KnowledgeMail (http://www.tacit.com)

[4] Kautz, H., Selman, B., Shah., M., “Referral Web: combining social networks and collaborative filtering” in*Communications of the ACM*, Vol. 40, Issue 3, pp. 63-65, 1997.

[5] Ha-Thuc, V., Venkataraman, G., Rodriguez, M., Sinha, S., Sundaram, S., Guo, L. "Personalized Expertise Search at LinkedIn", 2016.

[6] https://maj.io/#/

[7] http://yagajobs.co.uk

[8] Seid, D., Kobsa, A. "Demoir: A hybrid architecture for expertise modelling and recommender systems". 2000.

[9] Campbell, C.S., "Expertise identification using Bibliography 189 email communications", 12th Int. Conf. on Inform. and Knowl. Manag. ACM Press. 2003.

[10] D'Amore, R. "Expertise community detection", 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press. 2004.

[11] Mockus, A., Herbsleb, J.D."Expertise browser: a quantitative approach to identifying expertise", 24th Int. Conf. on Software Engineering. ACM Press. 2002.

[12] Hawking, D."Challenges in enterprise search", 15th Australasian Database Conference. Australian Computer Society, Inc. 2004.

[13] Tsiporkova, E., Tourwé, T."Tool support for technology scouting using online sources" in *Springer*pp. 371–376. 2011.

[14] Singh, H."Developing a Biomedical Expert Finding System Using Medical Subject Headings" in*Healthcare Informatics Research*Vol. 19, Issue 4,pp. 243-249. 2013.

[15] Hristoskova, A."A Graph-based Disambiguation Approach for Construction of an Expert Repository from Public Online Sources", 5th IEEE Int. Conf. on Agents and Artificial Intelligence. 2013.

[16] Abramowicz, W."Semantically Enabled Experts Finding System - Ontologies, Reasoning Approach and Web Interface Design" in *ADBIS*Vol. 2, pp. 157-166. 2011.

[17] Bozzon, A."Choosing the Right Crowd: Expert Finding in Social Networks", EDBT/ICDT'13. Genoa, Italy. 2013.

[18] Craswell, N. "Overview of the TREC-2005 Enterprise Track", 14th Text Retrieval Conference. 2006.

[19] Balog, K."People search in the enterprise". PhD thesis, Amsterdam University. 2008.

[20] Toutanova, K. "Enriching the knowledge sources used in a maximum entropy part of speech tagger", the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. EMNLP/VLC-2000. 2000.

[21] Buelens, S., Putman, M., "Identifying experts trough a framework for knowledge extraction from public online sources". Master thesis, Gent University, Belgium, 2011.

[22] Boeva, V., Krusheva, M., Tsiporkova, E. "Measuring Expertise Similarity in Expert Networks", Proceedings of the 6th IEEE Int. Conf. on Intelligent Systems, pp. 53-57, 2012.

[23] Boeva, V., et al. "Data-driven Techniques for Expert Finding", ICAART 9th International Conference on Agents and Artificial Intelligence, pp. 535-542, Porto, 2017.

[24] Stankoff, D., Kruskal, J. "Time warps, string edits, and macromolecules: the theory and practice of sequence comparison", Addison Wesley Reading Mass. 1983.

[25] Sakoe, H. and Chiba, S. "Dynamic programming algorithm optimization for spoken word recognition". In IEEE Trans. On Acoust, Speech, and Signal Proc., ASSP-26, pp. 43-49, 1978.

[26] Cameron, D, L. "SEMEF: A Taxonomy-based Discovery of Experts, Expertise and Collaboration Networks". MS thesis, The University of Georgia. 2007.

[27] Hecht, F. "The Journal Impact Factor: A Misnamed, Misleading, Misused Measure" in*Cancer GenetCytogenet*, Vol. 4, pp. 77-81, Elsevier Science Inc. 1998.

[28] Seglen, P. O. "Why the impact factor of journals should not be used for evaluating research" in *BMJ*. Vol. 314, Issue 7079, pp. 497. 1997.

[29] Hirch, J. E. "An index to quantify an individual's scientific research output" in*PNAS* Vol. 102, Issue 46, pp. 16569-16572. 2005.

[30] Afzal, M.T., Maurer, H. "Expertise Recommender System for Scientific Community" in *Journal of Universal Computer Science* Vol. 17, Issue 11, pp. 1529-1549. 2011.

[31] Fellbaum, C., "WordNet: An Electronic Lexical Database". MIT Press, Cambridge. 2001.

[32] Miller, G. A. "WordNet: A lexical Database for English" in*Communications of the ACM* Vol. 38, Issue 11, pp. 39-41. 1995.

[33] Boeva, V. et al. "Measuring Expertise Similarity in Expert Networks", In 6th IEEE Int. Conf. on Intelligent Systems, IS 2012 IEEE Sofia Bulgaria, pp. 53-57. 2012.

[34] Boeva, V. et al. "Semantic-aware Expert Partitioning" Artificial Intelligence: Methodology, Systems, and Applicationsin *LNAI. Springer* Int. Pub. Switzerland. 2014.

[35] B. Ganter, B., Stumme, G. and Wille, R. Formal Concept Analysis: Foundations and Applications, LNAI, no. 3626, Springer-Verlag, 2005.

[36] Zhou, J., Shui, Y. "The MeSHSim package".

[37] Paliwal, K.K. et al. "A modification over Sakoe and Chiba's Dynamic Time Warping Algorithm for Isolated Word Recognition" in *Signal Proc* Vol. 4, pp. 329-333. 1983.