Evolvable Semantic Platform for Facilitating Knowledge Exchange

Nikolay Karpov, Eduard Babkin, and Alexander Demidovskij

National Research University Higher School of Economics,
25/12 Bolshaja Pecherskaja str.,
Nizhny Novgorod 603155,
Russia
{nkarpov,eababkin,ademidovskiy}@hse.ru

Abstract. The authors propose new formal foundations and design approach to develop an evolving semantic platform for finding experts relevant to events arising in the open environment of modern economical clusters. This work offers a new implementation of probabilistic latent topic modeling method with two linked indicators (categories and experts) to mach expertise. In order to show feasibility of the solution a distributed and service-oriented software prototype of the web-based semantic platform was developed. Solution provides results with high precision scores and evolves in accordance with changes over time. Fusing together ontology-aided expertise matching and service-oriented software design suitable for developing evolvable semantic applications our approach facilitates effective and efficient knowledge exchange. That prototype called EXPERTIZE was evaluated for a particular case of experts finding in the university clusters.

1 Motivation

The environment of a modern enterprise is rapidly changing under the influence of a number of factors: socio-cultural, economical, political, technological and other. In this situation, inter-organizational exchange of expertise and professional knowledge plays an important role for sustainable operation of innovation enterprises. The emerging and successful growth of new forms of inter-organizational cooperation known as regional, innovative or university clusters in national economies is becoming a significant phenomenon of the modern world-wide socio-economical system [1]. Many researchers try to improve the learning performance of clusters. For instance, in a study by [2], the authors analyze the structures and relationship characteristics of major innovation clusters, and how the structural aspects and learning openness of these networks influence the performance of organizations in clusters. A study by [3] investigates the quantitative relationship between knowledge sharing, innovation and performance. Explicit knowledge sharing is considered to have more significant effects on innovation speed and financial performance, while tacit knowledge sharing has more significant effects on innovation quality and operational performance.

A university undoubtedly should be a catalyst for exchanging expertise and professional knowledge. Critical problems and major strategic choices should be commented on, discussed and exposed to multiple stakeholders including industry mass-media and society.

We believe that advanced methods of automated and automatic knowledge management belong to critical scientific foundations of modernizing the paradigm of information and knowledge exchange. A specifically designed combination of automated text processing and ontology-based knowledge engineering may improve the quality of information analysis and reduce the university's response time.

The paper contributes to the solution of the specified problem providing a systematic analysis of the expert retrieval task using topic modeling. We offer a new application of probabilistic topic models and suggest a service-oriented design of an evolvable semantic platform for finding university experts relevant to events arising in the economical clusters. In our research, evolvability of the semantic platform means that the expert retrieval solution adapts to changes of experts' interests and events in the open environment over time.

The novel application of the topic modeling approach consists in matching events with the semantics of two indicators: experts and categories using the Bayesian formula. There is two-fold linking between the indicators: first by semantics of the textual content, second by explicit tagging. This allows the user to speed up the process of expert finding by filtering by category. As a result, the proposed two-indicator approach provides university employees with the most recent and critical information about events, based on document analysis and the matching algorithm.

This paper is structured as follows. The second section contains a short overview of knowledge structures used in search and recommender systems. In Section 3, we describe a novel way of using probabilistic topic models for ontology-aided expertise matching. In Section 4, we provide a detailed software design of a system called EXPERTIZE, which uses probabilistic topic modeling for the expert finding task. Section 5 contains the results of the evaluation performed. Conclusions and a discussion of the implications of this work are provided in Section 6.

2 Related works

2.1 Semantic modeling

Considering the rapid growth of information generated by users of the Internet, it is essential to make connections in such a way that it is possible for everyone to gain the relevant information. It is the problem of constructing a semantic web. A set of core principles of the semantic web were developed in [4]. The idea of the semantic web is that the knowledge sharing process has to be effective, so that the community can get benefits from better integration. Some useful suggestions about the reuse of knowledge patterns were made by [5].

The inter-organizational knowledge management has also been recognized as a critical factor for an organization to remain competitive. This idea was reflected in the Media Information Logistics project (Media-ILOG). The goal of the Media-ILOG [6] was to improve the information flow inside a local newspaper JonkopingsPosten.

Another inter-organizational semantic application was proposed by [7]. The recent proposal included a specialization of the generic paradigm of ontological engineering, specific types of machine-readable RDF ontologies and an application of the temporal look at information relevant for team formation. In the software prototype InfoPort, aimed at solving the research team formation problem, the authors proposed to translate a user-specified query to a corresponding SPARQL query, which was then evaluated against a specific set of RDF repositories. The query result consisted of relevant categories of scientific classification taxonomies and keywords. The search algorithm of the InfoPort system retrieved everyone who was labeled with this query.

2.2 Expert finding

Some systems that cover several aspects of expert finding have already been developed. Initially, there were manual systems like SpuD [8], Sage People Finder [8]. Two of the earliest intellectual platforms dedicated to this problem were MITRE [9] and Panoptic by [10]. Nowadays, systems are much more complicated and precise; some examples are ArnetMiner [11], INDURE, Microsoft Academic Search [12], K-net[13], and SmallBlue [14].

The problem of finding the relevant and high-quality piece of advice is an extremely vital question nowadays. This search is about finding people who have relevant experience, or who are referred to as experts. There are more specific tasks derived from expertise retrieval: enterprise document search, learning to rank, entity retrieval, etc. As far as expert finding is concerned, we usually consider following approaches: profile-based, citation graph-based, textual contents-based, based on voting, and the hybrid one. The fundamental principle of the most approaches was systematized by Balog et al., [12]. Textual contents approach is widely discussed in the scientific world and deserves much more attention than it is currently paid. The core ideas of the modern document-based approach are described in the study of Fang and Zhai [15]. These scientists applied it to a language-based framework in an attempt to develop a better expert finding solution. However, the hybride approach is considered to outperform the profile-based and document-based ones [16].

As it is crucial to compare the different approaches to the expert finding task, some test collections have been developed. The most commonly used one is from the Enterprise Track at the Text Retrieval Conference (TREC 2005-2008)[17],[18], [19], [20] and corresponding benchmarks. This dataset consists of a collection of web pages and mainly textual features could be extracted. DBLP Computer Science Bibliography Dataset, contains the authors' publication records, is very rich on citation links (which enable the exploration of graph structures) and contains the publications' titles and abstracts [21].

Rich heterogeneous information derived from the textual contents, from the graph structure of the citation patterns for the community of experts, and from profile information about the academic experts can be combine multiple estimators of expertise based on a multisensor data fusion framework together with the Dempster-Shafer theory of evidence and Shannon's entrop [22]. However, there are still some obstacles such as using specific components in various methods, training requirements and others, which make the process of comparing these approaches extremely difficult.

2.3 Topic modeling

It is impossible to exaggerate the importance of topic modeling as the basic approach in the expert finding tasks. Initially, the Latent Dirichlet Allocation (LDA) was introduced by [23] as an extension of the Probabilistic Latent Semantic Analysis (pLSA) that includes prior distributions on the generation of topics and words. The topic mixture weights are not individually calculated for each document, but are treated as a k-parameter hidden random variable, where k is the number of topics. These variables that represent the word-topic and topic document distribution is a Dirichlet ones. Later, the idea of author-modeling was broadly developed by a range of leading scholars: [24] etc.

Academic knowledge discovery was also studied by [25], where the authors applied the group level topic modeling based on the LDA for the expert finding task.

The topic modeling approach has also been applied to analyzing any type of media resources. For instance, [26] proposed a complete system that retrieved experts from microblog platforms and ranked them according to their relevance. More importantly, a new Microblog LDA was presented, and according to the published results, it outperforms LDA for this specific task. In a study by [27], a new Trend-Sensitive LDA model was built for retrieving temporal trends in Twitter. Interestingly, there is also a study presenting a music recommender system [28] using Probabilistic Topic Modeling. There are also some papers on recommender systems for e-commerce, e.g. [29]. They proposed a dual approach: to recommend goods to clients and recommend item descriptions to sellers.

Based on LDA topic modeling,[16] developed an approach to solving the expert finding task, which outperforms the previous results on the TREC 2005-06 benchmarks. The indirect definition of several topics can build a more detailed and precise picture of personal expertise, so it can make the expert retrieval process more effective. The main principle was that using the LDA, it is possible to extract latent topics from a given text, and then one can find the distribution of possible expertise among all authors. This method seems to be a considerable milestone in solving the expert finding task.

3 Software Design of EXPERTISE

The described method for expert finding was implemented in a system for matching between relevant university experts and actual information events arising in

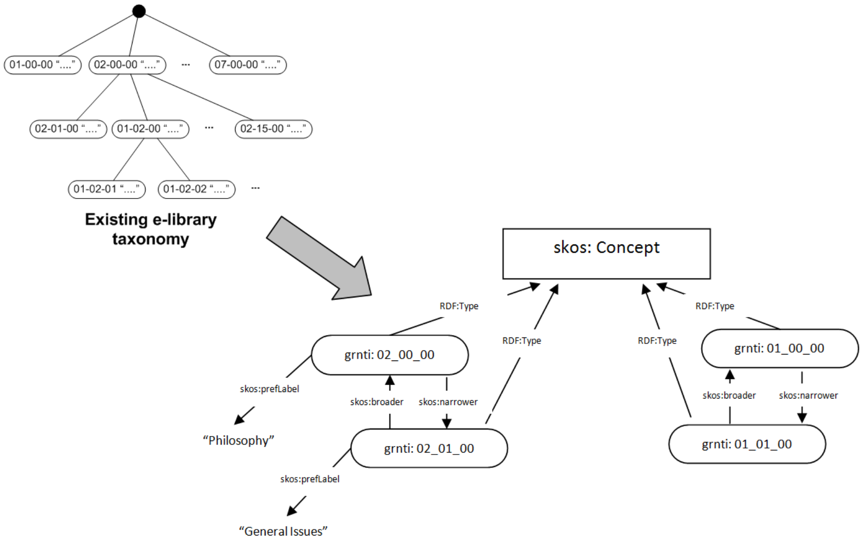

Fig. 1. The principles of mapping between the taxonomy and the ontology of scientific areas.

the open environment of the economical cluster. This system was called EXPERTIZE. The following services comprise the high-level design of the system (Figure 1):

1. Web Crawler;
2. Data Modeler;
3. Data Store;
4. Graphical User Interface (GUI);
5. Matcher.

EXPERTIZE actively uses our InfoPort Service [7]. This semantic service provides information about more than three hundred employees of the Higher School of Economics (NRU HSE)[1] branch in Nizhny Novgorod in the form of a formal ontology. The InfoPort data is represented in RDF triples. The triples include hierarchical information as it originally is in the source. The first level is an alphabetically ordered list of a group of academics, the second is an academic with his/her personal interests and papers, and the third is papers with their features.

The components of the EXPERTIZE system can be divided into Online and Offline services. Both interact with the InfoPort via the native REST interface. The Offline service works within the period of a month to update information regularly. The Online service works on demand, when a user activates it via the web interface.

[1] http://www.hse.ru/en/

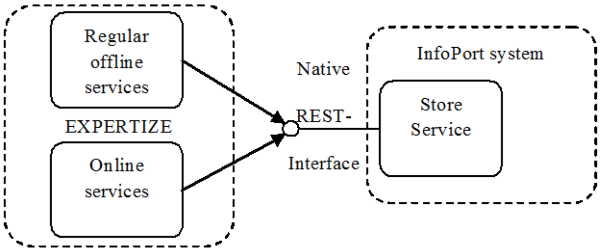

Fig. 2. Interaction of EXPERTIZE services with the InfoPort platform.

The offline processing begins with a crawler Service by a scheduler. The Crawler makes a request via the REST-interface to the InfoPort Store Service to get a list URI (Uniform Resource Identifiers) for papers. As soon as each paper is available online, the Crawler gets it by the URI and extracts its features from a page using an XML parser. The features of a paper include the authors, title, abstract, free keywords, and scientific categories based on the ontology. This information is collected in the Data Store with the help of a MySQL[2] base as a temporary raw data. The Crawler Service is implemented in the Python[3] programming language using the Lxml[4] library for HTML processing (Figure 3).

The preprocessing in the Data Modeler service includes the following steps:

- getting temporary raw data;
- tokenizing the text;
- lemmatizing the tokens;
- indexing the words using the dictionary of lemmas;
- filtering out the words that are too frequent (stop words) or too rare (used only once);
- indexing the authors and scientific categories;
- forming the bag of words using the lemmas, authors and categories;
- building the LDA model with a given number of topics K.

The online processing is performed on demand, when a user opens the Web GUI. The web interface activates the RSS Newsfeed, which gets and displays 10 last news posts from the RSS feed and an empty textbox. A user can choose one of the 10 news posts or paste the text into the textbox manually. When a user specifies an input query, the GUI transfers it to the Matcher. In turn, this component performs an online semantic search. A semantic representation of the event is matched with semantic representations of scientific categories and experts by applying the formula (8) and (9) and selecting top 5 of the units. Then, the Matcher component returns the 10 URIs to the GUI. To provide user

[2] http://www.mysql.com
[3] http://www.python.org
[4] http://lxml.de

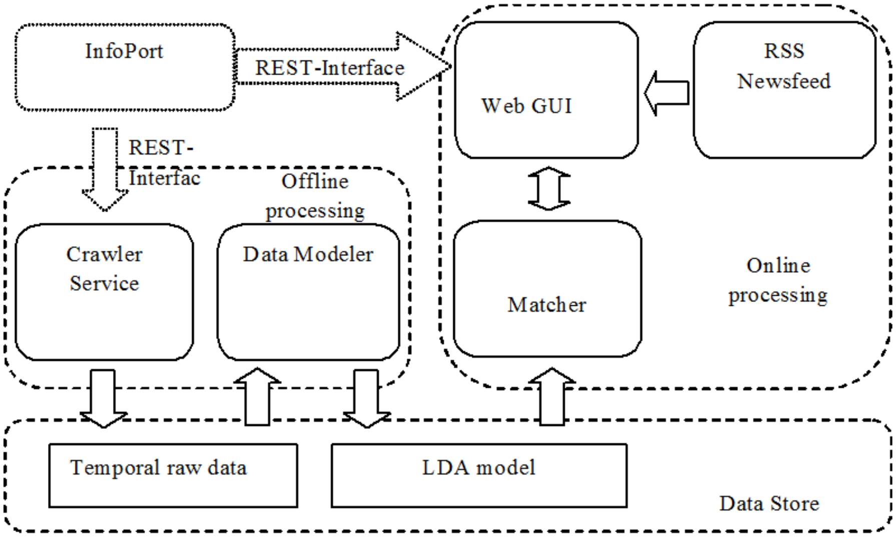

Fig. 3. Design of the EXPERTIZE system.

friendly output of the search result, the GUI component makes a request to the Infoport Service. It gets features of the selected units: the expert's full name, photo URL, and position.

4 Evaluation

We used information about more than three hundred experts from the Higher School of Economics (NRU HSE)[5] branch in Nizhny Novgorod, as it was mentioned above. The number of texts created by the experts and used to build the LDA model is 4132, all written in the Russian language.

The expert matching algorithm was evaluated using mean average precision of top k extracted most probable experts and categories. As input, we used our own test set that consists of 250 texts in Russian on the following topics: linguistics, law, politics, economics, management, mathematics and informatics. For each text, EXPERTIZE recommended the 10 experts and 10 categories with highest probabilities. The probability score calculated using formulas (6) with equiprobable and not equiprobable distribution of $P(T_i)$. For equiprobable case the probability was constant. For not equiprobable case the probability was estimated as follow:

$$\hat{P}(T_i) = \frac{T_i}{sum_i(T_i)} \tag{1}$$

[5] http://www.hse.ru/en/

T_i is a number of documents in collection where term i exists, $sum_i(T_i)$ is a total number of documents.

Let be a set of correct terms (experts, categories) and be a set of terms among the top k terms with highest probability. Then precision, and mean average precision (MAP) at k are calculated as follows, respectively:

$$Precision(k) = \frac{|R \cap \hat{R}_k|}{|\hat{R}_k|} \tag{2}$$

$$MAP(k) = \frac{1}{k} \Sigma_{i=1}^{k} P(i) \tag{3}$$

The quality of a similarity measure is assessed with the following statistics $Precision(i), i = \overline{1, 10}$, and $MAP(10)$. Table 1 shows the performance of the EXPERTIZE platform for experts and categories separately. Mean average precision for top 10 experts is 0.57. It is higher than the results achieved in English TREC 2006 and 2005 datasets. At the same time, not equiprobable formula shows higher results then equiprobable ones. We explain this by the fact that

Table 1. Precision for top k of experts and categories.

Score	Equiprobable		Not equiprobable	
	Experts	Categories	Experts	Categories
Precision (10)	0.86	0.72	0.92	0.68
Precision (9)	0.82	0.65	0.89	0.66
Precision (8)	0.75	0.62	0.83	0.66
Precision (7)	0.65	0.55	0.78	0.63
Precision (6)	0.65	0.48	0.70	0.61
Precision (5)	0.62	0.44	0.63	0.58
Precision (4)	0.55	0.48	0.55	0.56
Precision (3)	0.44	0.41	0.46	0.53
Precision (2)	0.24	0.41	0.34	0.43
Precision (1)	0.17	0.37	0.19	0.20
MAP (10)	0.57	0.51	0.62	0.55
MAP English TREC 2006	0.471	-	-	-
MAP English TREC 2005	0.248	-	-	-

our dataset had a lower number of documents and it's distribution is non-uniform over different categories and experts.

To sum up, the experiment shows that the proposed approach makes accurate recommendations based on the text under consideration, which makes it a useful tool for expert finding and revealing latent topics.

5 Conclusion

Topic modeling is successfully used to enhance search and recommendation in enterprise software [30]. Our EXPERTIZE platform applies topic modeling to online expert recommendation using the university community as the expert pool. We realize and evaluate an algorithm for matching events with a semantic of two indicators: experts and categories using Bayesian formula. Two linked indicators for expert retrieval with high precision allows easy search of the relevant person. As a source of categories and keywords two taxonomies are used together as a machine-readable ontology of scientific areas. The first part of this ontology is the international UNESCO nomenclature for fields of science and technology. The second part of the ontology was developed by our research team based on the Russian scientific classification called e-library.

The EXPERTIZE platform is organized as an evolvable and service-oriented application. It provides results with high precision scores and evolves in accordance with changes over time. In our research, evolvability of the semantic platform means that the expert retrieval solution adapts to changes of experts' interests and events in the open environment. We believe that the results of our study have a high potential in our global fast-changing world.

In the future, information used to retrieve relevant experts could be taken from more than one sensor and multiple estimators can be combined as proposed by Moreira & Wichert [22]. We believe that information derived from textual content could also be improved by exploiting state-of-the-art topic modeling instead of the LDA, but it should be evaluated on the benchmarks first.

6 Acknowledgment

The reported study was funded by RFBR according to the research Project No. 16-06-00184 A.

References

1. B. Asheim, P. Cooke, and R. Martin, "Clusters and regional development: Critical reflections and explorations," *Economic Geography*, vol. 84, no. 1, pp. 109–112, 2008.
2. J. Choi, A. S. Hyun, and M.-S. Cha, "The effects of network characteristics on performance of innovation clusters," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4511–4518, 2013.

3. Z. Wang and N. Wang, "Knowledge sharing, innovation and firm performance," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8899–8908, 2012.
4. A. Ohgren and K. Sandkuhl, "Towards a methodology for ontology development in small and medium-sized enterprises." in *IADIS AC*, 2005, pp. 369–376.
5. K. Hammar, F. Lin, and V. Tarasov, "Information Reuse and Interoperability with Ontology Patterns and Linked Data," in *Business Information Systems Workshops.* Springer, 2010, pp. 168–179.
6. K. Sandkuhl, A. Ohgren, A. Smirnov, N. Shilov, and A. Kashevnik, "Ontology construction in practice: Experiences and recommendations from industrial cases," in *9th International Conference on Enterprise Information Systems, 12-16, June 2007, Funchal, Madeira–Portugal*, 2007.
7. E. Babkin, N. Karpov, and O. Kozyrev, "Towards Creating an Evolvable Semantic Platform for Formation of Research Teams," in *Perspectives in Business Informatics Research*, ser. Lecture Notes in Business Information Processing, A. Kobyliński and A. Sobczak, Eds. Springer Berlin Heidelberg, Jan. 2013, no. 158, pp. 200–213.
8. I. Becerra-Fernandez, "Searching for experts on the Web: A review of contemporary expertise locator systems," *ACM Transactions on Internet Technology (TOIT)*, vol. 6, no. 4, pp. 333–355, 2006.
9. D. Mattox, M. T. Maybury, and D. Morey, "Enterprise expert and knowledge discovery." in *HCI (2)*, 1999, pp. 303–307.
10. N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins, "P@ noptic expert: Searching for experts not just for documents," in *Ausweb Poster Proceedings, Queensland, Australia*, 2001.
11. D. Kempe, J. Kleinberg, and v. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2003, pp. 137–146.
12. K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, L. Si *et al.*, "Expertise Retrieval," *Foundations and Trends in Information Retrieval*, vol. 6, no. 2-3, pp. 127–256, 2012.
13. N. S. Shami, Y. C. Yuan, D. Cosley, L. Xia, and G. Gay, "That's what friends are for: facilitating'who knows what'across group boundaries," in *Proceedings of the 2007 international ACM conference on Supporting group work.* ACM, 2007, pp. 379–382.
14. K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: combining text and social network analysis," in *Proceedings of the 2007 international ACM conference on Supporting group work.* ACM, 2007, pp. 117–126.
15. H. Fang and C. Zhai, "Probabilistic models for expert finding," in *Advances in Information Retrieval.* Springer, 2007, pp. 418–430.
16. S. Momtazi and F. Naumann, "Topic modeling for expert finding using latent Dirichlet allocation," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 5, pp. 346–353, 2013.
17. N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC 2005 Enterprise Track." in *Trec*, vol. 5, 2005, pp. 199–205.
18. I. Soboroff, A. P. de Vries, and N. Craswell, "Overview of the TREC 2006 Enterprise Track." in *Trec*, 2006.
19. P. Bailey, A. P. De Vries, N. Craswell, and I. Soboroff, "Overview of the TREC 2007 Enterprise Track." in *TREC*, 2007.
20. K. Balog, P. Thomas, N. Craswell, I. Soboroff, P. Bailey, and A. P. De Vries, "Overview of the TREC 2008 enterprise track," DTIC Document, Tech. Rep., 2008.

21. Z. Yang, J. Tang, B. Wang, J. Guo, J. Li, and S. Chen, "Expert2bole: From expert finding to bole search," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(KDD'09)*, 2009, pp. 1–4.
22. C. Moreira and A. Wichert, "Finding academic experts on a multisensor approach using Shannon's entropy," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5740–5754, 2013.
23. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, no. 3, pp. 993–1022, 2003. [Online]. Available: http://jmlr.org/papers/volume3/blei03a/blei03a.pdf
24. T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.
25. A. Daud and F. Muhammad, "Group topic modeling for academic knowledge discovery," *Applied Intelligence*, vol. 36, no. 4, pp. 870–886, 2012.
26. S. Xianlei, Z. Chunhong, and J. Yang, "Finding Domain Experts in Microblogs," 2014.
27. M.-C. Yang and H.-C. Rim, "Identifying Interesting Twitter Contents Using Topical Analysis," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4330–4336, Jul. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2013.12.051
28. Z. Hyung, K. Lee, and K. Lee, "Music recommendation using text analysis on song requests to radio stations," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2608–2618, 2014.
29. K. Christidis and G. Mentzas, "A topic-based recommender system for electronic marketplace platforms," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4370–4379, 2013.
30. K. Christidis, G. Mentzas, and D. Apostolou, "Using latent topics to enhance search and recommendation in Enterprise Social Software," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9297–9307, 2012.