A Survey of Intelligent Question Routing Systems

Bojan Furlan, Bosko Nikolic, Member of the IEEE, and Veljko Milutinovic, Fellow of the IEEE
School of Electrical Engineering, University of Belgrade, Serbia
{bojan.furlan, bosko.nikolic, veljko.milutinovic}@etf.bg.ac.rs

Abstract - This paper represents a survey of the existing research in the domain of intelligent question routing. The survey starts from an original presentation paradigm that generalizes the essence of approaches found in the open literature. The presentation paradigm includes three basic processing stages related to the three major problems of system implementation. Various research efforts use different approaches for implementation of each one of the basic processing stages. Each particular approach is presented here using the same template. All these approaches are enlisted, discussed, and presented using a table, for easier comparison. The outcome of this analysis is a proposal for a new approach based on a generalized treatment of the user knowledge profiling. Major contributions of this survey paper are: (a) original presentation paradigm, (b) detailed description of existing approaches, (c) comparative study of existing approaches, and (d) proposal of a new approach to user knowledge profiling, which enables uniform incorporation of new information sources in the form of software agents.

I. INTRODUCTION

Multidisciplinarity and collaboration are essential drivers for innovation, thus intelligent question routing systems (IQRS) aim to serve as a knowledge exchange medium in an arbitrary field of expertise. The goal of this survey paper is to shed light on a selected set of novel approaches, which is of importance for a number of applications, where intensive communication between users is required (e.g., large enterprises, e-government agencies, health care system, army, etc.) Other applications can involve a support in educational and collaboration processes, where IQRS facilitates an efficient and effective knowledge exchange between scientists, researchers, university staff, and students. The benefit coming from deployment of such systems includes: (a) reducing stress on experts, which are a valuable resource and (b) increasing the system owners' (enterprise, government, university) quality of service.

This survey starts from a discussion on how the IQRS domain of research is related to other similar fields and why it is important. Then, it gives an original presentation paradigm that generalizes the essence of approaches found in the open literature. The presentation paradigm includes three basic processing stages. Each particular approach is presented using the same template. All these approaches are enlisted, discussed, and presented using a comparative table, for easier comparison. The outcome of this analysis is a proposal for a new approach based on a generalized treatment of the user knowledge profiling.

II. BACKGROUND

Question Answering (Q/A) is the task of automatic answering a question posed in a natural language. The main purpose of Q/A systems was, and still is, to move the retrieval focus from Document Retrieval to Information Retrieval, by extracting relevant and concise answers to a wide range of open domain questions. For finding the answer, Q/A systems use diverse data sources from pre-structured databases to large collections of documents written in a natural language (text corpus). This text corpus can consist of formal documents like compiled newswire reports [1], or from noisier ones (and not strictly formatted) such as blogs from the World Wide Web [2]. However, as stated before, one of the key characteristics of today's scientific research is multidisciplinarity. Sometimes it is necessary to adopt fundamental knowledge from a variety of domains. On the contrary, quite many experts with needed knowledge exist within some institution or university. For a young researcher or student it would be very helpful to contact directly a person competent in a particular domain and ask him/her for an advice or instruction. The efficiency of finding the right person can be gained by using a software system for intelligent question routing.

Also, an extensive research has been conducted in the field of semantic query routing (SQR) in peer-to-peer networks. One example of how to locate peers that are relevant with respect to a given query is by building a semantic overlay network [3]. Queries are routed through a super-peer where every peer needs to explicitly advertise its content. Another example is implicit content identification based on social metaphors [4]. With advancement of text processing tools and a recent boom of social networks, the synergy of Q/A and SQR has become possible, so question routing between users - IQRS is an open research issue. We refer to questions as a free-form text, as opposed to structured or semi-structured queries. Accordingly, the rest of the paper represents a survey of related work focused only on IQRS and related problems of importance for the paradigm introduced in this paper.

III. GENERALIZATION OF SELECTED APPROACHES

The viewpoint of this survey is best represented by the notions of Figure 1. Authors of this survey assume that question routing is a complex process influenced by both static and dynamic parameters, so the results of the presented research are more widely applicable.

978-1-4673-2278-2/12/$31.00 ©2012 IEEE

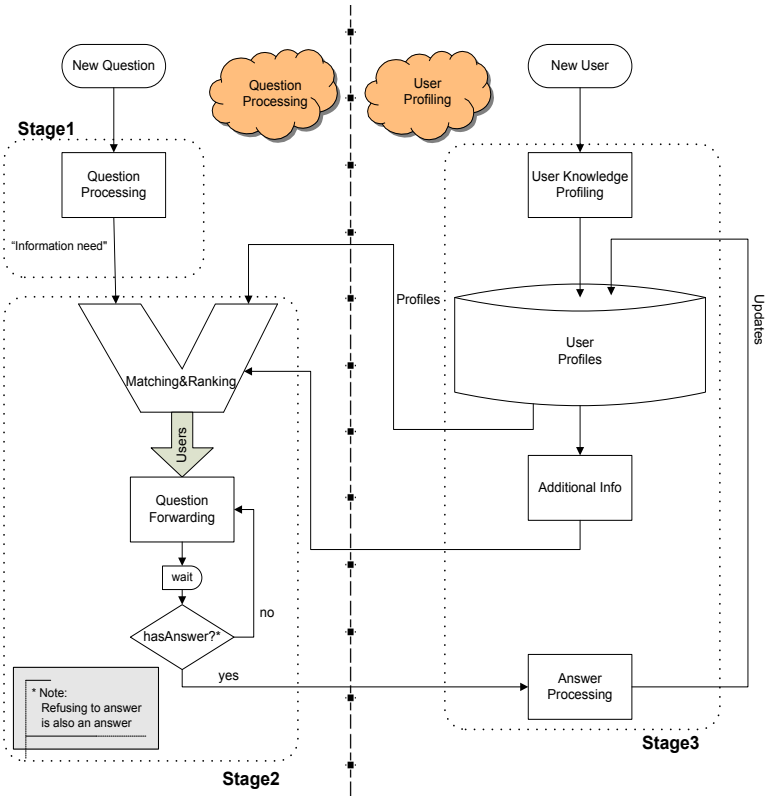

Figure 1. An Original Presentation Paradigm: Generalization of Approaches from the Open Literature

Structure of this process is divided in two parts, which simultaneously process: (a) new questions (Question Processing) and (b) new or existing users (User Profiling). Both parts consist of stages represented in Figure 1. Each stage contains one or more modules, which are implemented using an algorithm from a relatively large pool of algorithms. Stage1 contains *Question Processing* module, Stage2 contains *Matching&Ranking* and *Question Forwarding* modules, and Stage3 contains following modules: *User Knowledge Profiling*, *Additional Info*, and *Answer Processing,* as well as *User Profiles* repository.

Each stage is related to the three main issues that are defined by three questions elaborated in the text to follow:

Question#1: *How to identify information need from a question?*

In IQRS, the requirement for a question analysis is only to be able to understand the question sufficiently for routing it to a competent answerer. This is a considerably simpler task than the challenge facing an ideal Q/A system, which must attempt to determine exactly what piece of information the user is seeking (e.g., to translate information need into search keywords), to evaluate whether a founded content includes that piece of information, and to extract it to a human-understandable format. By contrast, in IQRS, it is the human answerer who has the responsibility for determining the relevance of an answer to a question, and that is a function which human intelligence is well-suited to perform.

The task of question analysis and information need extraction is presented by the *Question Processing* module in Figure 1. The output of this module is in a form of identified topics or terms, which are forwarded to the *Matching&Ranking* module.

Criteria of interest for **Question#1**:

1) User interaction type:
   a) With question annotation (e.g., tags, categories),
   b) Without question annotation.
2) Algorithm extraction type:
   a) Natural Language Processing techniques - NLP (e.g., stemming, Part Of Speech - POS processing and filtering, synonym lookup),
   b) Data Mining/Machine Learning techniques - DM (e.g., trained topic classifiers) or ML (topic modeling).

Possible improvement avenues for **Question#1:** Question visualization.

Question#2: *How to find competent users for a particular question?*

Given the information about the question derived from the question-processing module, the task of finding competent users is performed by matching the recognized information need from the question against the available knowledge profiles, and ranking them in an ordered list of users (or "candidate answerers") who should be contacted to answer the question. This matching can be realized as exact matching or by computing a semantic similarity.

As shown in Figure 1, the inputs of the *Matching&Ranking* module is an information need recognized form the question, available profiles from the user profiles repository, and additional info like availability, referral rank, etc. The output is an ordered list of users that is submitted to the *Question Forwarding* module.

Criteria of interest for **Question#2**:

1) Model organization (centralized, distributed),
2) Semantic matching (with or without semantic similarity matching between a user profile and an information need recognized from a question).

Possible improvement avenues for **Question#3:** Semantic and string similarity incorporation.

Question#3: *How to accurately profile user's knowledge from various information sources?*

Knowledge can be classified broadly as either explicit or tacit [5, 6]. Explicit knowledge consists of facts, rules, relationships, and policies that can be faithfully codified in paper or electronic form. Since it is explicitly expressed, it can be shared without a need for discussion. By contrast, tacit knowledge (or intuition) requires interaction. This kind of knowledge underlines personal skill, and is largely influenced by beliefs, perspectives, and values. Its transfer requires face-to-face contact or even apprenticeship. Since individual knowledge is learned (internalized) into the human brain, the psychological approach by observing the subject's characteristics from the performed behavior has to be applied. In this case, the observed behavior is represented by the content that a user generates. This content maps to explicit and tacit classification of knowledge to some extent, where explicit knowledge is mostly expressed within the published documents, such as scientific papers, books, articles or blogs, while email communication and content generated during the question-answering process can identify the tacit knowledge. As a result, both sorts of information are valuable for profiling of user knowledge.

The IQRS keeps user profiles in a repository that is constantly being updated. Besides expertise that is prime information kept about the user, the profile repository can also contain some other information (not directly related to knowledge) like response rate or affiliation. As shown in Figure 1, for a new user, the initial profile is created in the *User Knowledge Profiling* module. Afterwards, updates are gathered from the question-answering process (e.g., correct or incorrect answers rates) or from some external updates (e.g., manually changing its profile).

Criteria of interest for **Question#3 -** user profiling methodology (by information source and expertise identification):

1) Text (posts on forums, blogs, emails, etc.):
   a) Natural Language Processing techniques - NLP (e.g., stemming, ad-hoc named entity extractor),
   b) Data Mining/Machine Learning techniques - DM (e.g., classification, clustering) or ML (topic modeling),
   c) Recommender System (RS) model.
2) Other (social network linkage graph, response rate, etc.)
   a) ad-hoc model,
   b) Recommender System (RS) model.

Possible improvement avenues for **Question#1**: Profile integration.

IV. Presentation of Selected Approaches

Approaches presented in the text to follow address the three main issues that define IQRS in a characteristic manner. For easier comparison, all approaches are also presented in Table1. Each column in the table corresponds to a particular element of the intelligent question routing process from Figure 1. Each table entry includes the name and a short description (within the criteria of interest). Also, all analyzed approaches are described in a similar way, including the information according to the following template: 7Ws (who, when, what, etc.) or a subset there of, essence, structure, relevant details, applications, and pros & cons.

A. iLink

Davitz et al [7] in 2007 proposed a model for social search and message routing named iLink. They focused on the problem how to model social networks and how those networks accomplish tasks through peer-to-peer production style collaboration. The social network is represented as a graph with nodes and links. Expertise, response rates, and referral rates are maintained for every node (user). For Question Analysis iLink uses NLP, particularly stemming, synonym lookup, and stop word removals. Users are also allowed to tag questions in order to improve system's performance. For User Knowledge Profiling from text sources DM technique (clustering) is used. Other maintained parameter is response score, which is in function of response rate, response accuracy, etc. iLink model organization is centralized (as a supernode in the social network), but it can also be used in a decentralized manner. For Matching&Ranking it does not employ any semantic similarity between terms, but as an Additional Info it maintains a referral rank about the user, which correlates to popularity referrals from other users.

TABLE I. COMPARISON OF SELECTED APPROACHES

	1. Question Processing		2. Matching & Ranking		3. User Knowledge Profiling		4. Additional Info.
	Annotation	*Analysis*	*Model Organization*	*Semantic Matching*	*Text*	*Other*	
A. iLink	Tagging	NLP	Centralized (can be Distributed)	No	DM	Response Score	Referral Rank
B. PLSA in CQA	No	ML	Centralized	No	ML	No	No
C. Question Routing Framework	No	No	Centralized	No	RS model	No	Availability
D. Aardvark	Tagging	DM	Centralized	Yes	DM & NLP	RS Model	Connectedness, Availability
E. Yahoo! Answers Recommender System	Categories	NLP	Centralized	No	RS model	RS model	Group of user attributes
F. SQM	No	No	Distributed	No	No	Expertise Score	Response Rate

The iLink model has been used to develop a system for generation of Frequently Asked Questions (FAQ) in a social network - FAQtory. The system facilitates generation of a repository of question/answer threads, so when users send questions to the system they are presented with a list of related question/answer pairs, a list of experts on the topics found in the question, and as a last resort, search results from the web.

An interesting idea introduced with iLink, is that it allows incremental answering. At each step in a query thread, user nodes can contribute some information even if that information does not qualify as an answer. This information can be about the query itself or it can simply be some evidence about where knowledge might exist in the network (e.g., who knows something, who knows somebody). On the other hand, iLink does not use semantic similarity matching between extracted terms and incorporation of external information into user profiles is not trivial.

B. Probabilistic Latent Semantic Analysis in Community Question Answering Portals

Qu et al [8] in 2009 proposed a question recommendation technique using the Probabilistic Latent Semantic Analysis (PLSA) that helps users to locate interesting questions in Community Question Answering (CQA) portals such as Yahoo! Answers. For User Knowledge Profiling from text sources the PLSA topic modeling technique is used. There are no other parameters incorporated in the user knowledge profile. Also, for Question Analysis the same machine learning technique (PLSA topic modeling) is used. There is no possibility of question annotation and no Additional Info about user is maintained. Matching&Ranking is centralized and it is not using any semantic similarity match between extracted terms.

Despite the lack of many analyzed attributes, this paper is included in the survey since it introduced an innovative approach to knowledge profiling based on the topic modeling technique. Also, it proposed a novel metric to evaluate the approach performance by matching a recommended user's rank with the best answerer's rank in Yahoo! Answers dataset.

C. Question Routing Framework

Li and King [9] in 2010 proposed a framework called Question Routing (QR) that ranks the answerers in CQA. User Knowledge Profiling is done twofold: with and without consideration of an answer quality. The first estimation is modeling potential answer quality based on quality of previously answered questions by the user. The second one uses only term frequency for calculating similarity between a particular question and all previously answered questions. As an Additional Info, Availability is estimated. It is assumed that a user is available to provide answers for the routed questions when is logged on the system, so estimation is made by modeling this problem as a trend analysis problem in time-series data mining. Matching&Ranking is centralized and it calculates for each potential answerer the final QR score as a linear combination of estimated expertise score and availability score.

The QR framework considers both users' expertise and users' availabilities for providing answers in a range of time. However, it is hard to incorporate other parameters in users' profiles, there is no question analysis and annotation, and semantic matching between extracted terms is not incorporated.

D. Aardvark

Horowitz and Kamvar [10] in 2010 presented a commercial system named Aardvark. It represents a social search engine where users can ask a question, either by instant message, email, web input, or voice. Questions are analyzed with DM technique (a combination of trained topic classifiers), and user can additionally annotate them with tags. To find someone that is most likely to be able to answer a question, Aardvark routes this question to persons in the user's extended social network. Therefore, User Knowledge Profile incorporates its extended social network, which indexes affiliation and friendship information for every user

and their friends, representing a Friends-of-Friends social graph. User has an option of importing this information from existing social networks like Facebook, LinkedIn, or webmail contacts, or manually inviting friends to join. Simultaneously, for User Knowledge Profiling from text, Aardvark maintains the list of topics about which the user has some level of interest. These topics are identified from several sources: manually indicated by user or a friend that invites him/her, parsed from a profile page or an account on which he/she regularly posts status updates (e.g., Twitter or Facebook) and finally, observing the user's behavior on answering (or electing not to answer) questions about particular topics. For topic extraction a combination of DM & NLP is used, particularly support vector machine & ad-hoc named entity extractor. Also, attributes like Connectedness and Availability are maintained for every user as an Additional Info. Aardvark model organization is centralized and for Marching&Ranking of potential answerers both extended social network and topics extracted from text are used. Similarity between users is modeled employing recommender systems techniques with extended social network attributes, like demographic similarity, profile similarity, social connections, etc. Similarity between extracted topics from question and topics from user's profile is calculated using corpus-based semantic similarity, which is computed over Wikipedia and other text corpora.

As indicated by authors, the Aardvark search algorithm is being put on intimacy, where the user's trust in received answer is based on knowing the answerer (directly or indirectly from the social network). Thus, questions are routed primarily within user's extended social network. As reported, this provides results for questions that are in a context of user's social or demographic proximity (e.g., giving opinion about restaurant nearby or advice about dating). However, we want to emphasize another dimension of trust in the received answer, based on the answerer's knowledge reputation. This particularly stands for questions that are highly expert-oriented, where the user's information need possibly cannot be satisfied within its social network. Therefore, in this context, the more objective user's knowledge profiling is needed in order to effectively and efficiently forward questions to competent users.

E. Yahoo! Answers Recommender System

Dror et al [11] in 2011 also addressed a need for a mechanism in CQA portals (Yahoo! Answers in particular) that would expose users to questions they can relate to and possibly answer. Question Analysis is implemented using NLP techniques: stemming, stop word removals, and POS processing and filtering. Also, user has to annotate questions by assigning them to categories. System's architecture is centralized and Matching&Ranking is based on a multi-channel recommender system technology. To fuse and generalize information that represents multiple social and content signals from users, a single symmetric framework is constructed which incorporates and organizes these signals according to channels. Content signals are used for User Knowledge Profiling from text and they relate mostly to text attributes and categories of questions and associated answers. Other attributes are also included in a form of social signals, which capture the various user interactions with questions, such as asking, answering, voting, etc.

As authors claim, the key objective of this approach was to satisfy the sole asker in a variety of questions, some factoid but many being subjective where the notion of expertise is irrelevant. This differs from expert search task that tries to identify an authoritative answer that would satisfy most. Also, in a context of Yahoo! Answers system external sources of social relations between users are not available, so the main focus was to differentiate between various user-question interactions.

F. SQM

Banerjee and Basu [12] in 2008 proposed Social Query Model (SQM) for decentralized search, which has the Pagerank model and certain Markov Decision Processes as special cases. The social network is represented as a graph with nodes and links. The model does not consider question analysis and there is no question annotation. The organization is decentralized and Matching&Ranking is based on a distributed approximation algorithm, which computes optimal query routing policy. User's knowledge profile includes only expertise score and as an Additional Info response rate is incorporated. Therefore, in the context of the model this policy is simultaneously optimal for all nodes, in that no subset of nodes will jointly have any incentive to use a different local routing policy.

To some extent, all previously presented approaches are complementary to SQM, since the focus was not on query routing within nodes of social network, but on identifying a user's potential to give the correct answer and matching that potential to particular question. Therefore, the potential can be characterized by different factors, such as expertise and responsiveness, which are input parameters within SQM.

V. SUGGESTED APPROACHES FOR FUTURE RESEARCH

This section contains ideas of potential interest for future research.

A. Question visualization

Questions typically consist of a text, which is not too long, so one solution is that a question-processing module can be developed using NLP or DM/ML techniques. However, tools for automatic information extraction, in general, can be insufficiently precise and can omit some valuable information. Also, short questions are often ambiguous. Having that in mind, the most effective solution is an interactive user interface that allows communication between the question-processing module and a user that posed the question. This approach combines fully automatic text processing and manual correction of results, giving the user a possibility to increase the accuracy of the output. On the other hand, automated processing can produce more results that would be usually forgotten.

Figure 2. Question visualization: Example of a generated concept cloud

We propose that the discovered concepts should be visually represented in a form of a concept cloud (TagCloud visualization). One benefit of this approach is that "the more significant the concept is, the bigger the font size it has," which provides a more intuitive representation of relations between specific concepts and their importance in the question. Figure 2 represents an example of generated concept cloud. More details about possible implementation can be found in [13].

B. *Semantic and string similarity incorporation*

The measure of similarity between a question and a user profile can be realized by computing exact mach of recognized topics or terms or more precise by calculating their semantic similarity. Therefore, bag of words approach can be used that employs corpus-based or knowledge-based measures of word similarity [14]. For each word in a profile, the method should identify the highest match from the question and then combine it in the overall measure of semantic similarity. Islam and Inkpen [15] proposed an improvement of this similarity measure by incorporating a string matching algorithm with a corpus-based measure of semantic word similarity. Therefore, we indicate that a possible improvement can be found in this direction as illustrated in Figure 3. This method, besides the semantic word similarity measure, incorporates the string similarity measure, so it performs better with typos or different forms of infrequent proper nouns [16].

C. *Profile Integration*

Bayesian probability (used in the proposed solutions) has a firm theoretical foundation and it is widely used in trust management, at present. However, the Bayesian approach does not have an adequate expressiveness and it needs some artificial construction. For example, user A answered 100 questions about a topic *c* and the quality of the answers rated by other users was 0.5. Then, we consider another case that A did not answer any question about topic *c*. In both cases the evaluated trust of the Bayesian approach in A's knowledge about the topic *c* is p(trust)=0.5 and p(distrust)=0.5. Therefore, the Bayesian approach does not have the ability to distinguish these two cases [17].

One possible improvement can be found in the Dempster-Shafer theory (DST), a mathematical theory of evidence that is a generalization of the Bayesian probability.

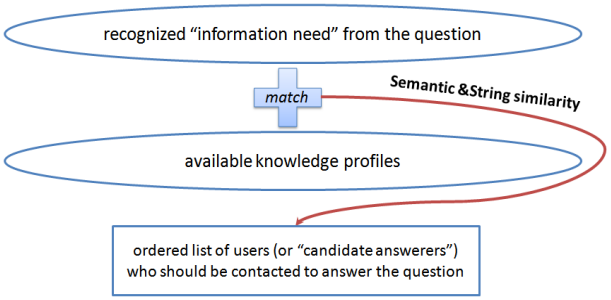

Figure 3. Semantic and string similarity incorporation: Ilustration of matching process

It can handle ignorance naturally and allows one to combine evidence from different sources. To arrive at a degree of belief (represented by a belief function) DST takes into account all available evidences. In addition, for profile integration within IQRS we suggest the use of an evidential trust model based on the Dezert-Smarandache theory [17], which is a generalization of DST, so it has a higher expressiveness.

VI. CONCLUSION

The objective of this paper was to systematically establish common characteristics of IQRSs, which are inherently heterogeneous, and to allow their uniform analysis. Hence, major contributions of this survey paper are: identification of the IQRS domain of research, generalization of approaches and the original presentation paradigm, description of selected approaches and their comparative study, and finally a proposal of three new directions for future research, which are related to the three main issues of importance for intelligent question routing. Our selection includes only the systems after 2007 (last 5 years) that recently generated most attention in the research community.

Findings and explanations of this survey are of interest to those how like to enter this emerging field of research, to understand the essential notions, and to obtain ideas for their future research. This may be of most benefit to PhD students.

Newly open problems fall into two basic categories: (a) to expand the survey to a wider body of knowledge and (b) to implement a prototype of the proposed new ideas and to evaluate their performance, comparatively with the best solutions from the open literature, or their approximated equivalents.

In conclusion, since questions and appropriate answers are the essence of IQRS, our attitude in this paper was: "*Prudens quaestio dimidium scientiae* - Half of science is asking the right questions" Aristotle (384 BC – 322 BC). Therefore, the three fundamental questions are asked and on their basis the presentation paradigm is built, which is supposed to be the main contribution of this paper.

ACKNOWLEDGMENT

Authors would like to thank Prof. Nenad Mitic for his advices and constructive discussions. The work presented

here was partially supported by the Serbian Ministry of Education and Science (projects III 44006, III 44009 and III 32047).

REFERENCES

[1] H. T. Dang, J. Lin, D. Kelly, and C. Hill, "Overview of the TREC 2006 Question Answering Track," in *TREC*, 2006.

[2] I. Ounis, C. Macdonald, and I. Soboroff, "Overview of the TREC-2008 Blog Track," in *TREC*, 2008.

[3] D. Faye, G. Nachouki, and P. Valduriez, "Semantic Query Routing in SenPeer , a P2P Data Management System," in *Network-Based Information Systems*, 2007, pp. 365-374.

[4] C. Tempich, D. Karlsruhe, S. Staab, and A. Wranik, "REMINDIN : Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors," in *WWW*, 2004.

[5] I. Rus and M. Lindvall, "Knowledge management in software engineering," *IEEE Software*, vol. 19, no. 3, pp. 26-38, May 2002.

[6] S. Frameworks, "Management of explicit and tacit knowledge," *Journal of the Royal Society of Medicine*, vol. 94, pp. 6-9, 2001.

[7] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A, "iLink: search and routing in social networks," in *WWW*, 2007.

[8] M. Qu, G. Qiu, X. He, and C. Zhang, "Probabilistic question recommendation for question answering communities," in *WWW*, pp. 1229-1230, 2009.

[9] B. Li, and I. King, "Routing questions to appropriate answerers in community question answering services," in *CIKM*, pp. 1585-1588, 2010.

[10] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *WWW*, 2010.

[11] G. Dror, Y. Koren, Y. Maarek, and I. Szpektor, "I want to answer; who has a question?: Yahoo! answers recommender system," in *KDD*, pp. 1109-1117, 2011.

[12] A. Banerjee and S. Basu, "A social query model for decentralized search," in *SNAKDD*, 2008.

[13] E. Varga, B. Furlan, and V. Milutinovic, "Document Filter Based on Extracted Concepts," *Transactions on Internet Research*, vol. 6, no. 1, pp. 5-9, January 2010.

[14] R. Mihalcea, C. Corley and C. Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," in *National Conference on Artificial Intelligence*, vol. 21, no. 1, pp. 775-780, 2006.

[15] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1-25, Jul. 2008.

[16] B. Furlan, V. Sivački, D. Jovanović, and B. Nikolić, "Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts," *Journal of Information Technology and Applications*, vol. 1, no. 1, pp. 65-71, 2011.

[17] J. Wang and H.-J. Sun, "A new evidential trust model for open communities," *Computer Standards & Interfaces*, vol. 31, no. 5, pp. 994-1001, Sep. 2009.