The role of semantic similarity for Intelligent Question Routing

Bojan Furlan
University of Belgrade
Bulevar kralja Aleksandra 73
RS-11120 Belgrade
Email: bojan.furlan@etf.bg.ac.rs

Slavko Žitnik
University of Ljubljana
Tržaška cesta 25
SI-1000 Ljubljana
Email: slavko.zitnik@fri.uni-lj.si

Boško Nikolič
University of Belgrade
Bulevar kralja Aleksandra 73
RS-11120 Belgrade
Email: bosko.nikolic@etf.bg.ac.rs

Marko Bajec
University of Ljubljana
Tržaška cesta 25
SI-1000 Ljubljana
Email: marko.bajec@fri.uni-lj.si

***Abstract*—Intelligent Question Routing Systems (IQRS) aim to serve as a knowledge exchange medium in an arbitrary field of expertise, where intensive communication between users is required (e.g., large enterprises, e-government agencies, technical support, health care system, army). Other applications can involve a support in educational and collaboration processes, where IQRS facilitates an efficient and effective knowledge exchange between scholars. The benefit coming from deployment of such systems includes: (a) reducing unnecessary "pinging" of experts, which are a valuable resource and (b) increasing the system owners' (enterprise, government, university) quality of service, since users are more satisfied with answers, because their questions are answered by the right persons. In this paper we investigate the role of semantic similarity for each stage of IQRS process. For question analysis we use semantic enrichment, more precisely semantic query expansion with ConceptNet, WordNet (Antelope), and SemNet. Also, for question routing stage we used techniques developed for semantic similarity between two paragraphs sentences. Finally, for evaluation we used special subsets of Yahoo! L6 Answers dataset from which we extracted three different types of users: (1) Top questioneers, (2) Top answerers, and (3) Top combination of previous user types to model interest and expertise. Over these users we build special semantic profiles and match them to questions, answers or the whole question threads, according to the specific IQRS stage.**

I. Introduction

TODOdefine that post can be answer or question

This demo file is intended to serve as a "starter file" for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.7 and later. I wish you the best of success.

mds

January 11, 2007

II. Related work

SemSim & AvailableTech

III. Semantic Similarity

IV. Algorithm

Max Sem sim **TODO**define what is evidence! **TODO**include link to the software repository

V. UserProfiling

TODOwhat is profiling, how we do profiling, what do we think of it ...

A. Sources

To build efficient user profiles, we collect a lot of evidences from user's posts. We separate evidences by source, which can origin from question body, answer body, question title, whole post thread or IQRS system. We further categorize evidences by their type, which can be one the following:

- **ConceptNet** [1] is a semantic knowledge base that describes general human knowledge. It includes words and common phrases from many written texts. They are related through open domain predicates and through common knowledge. The database was created manually and partially automatically from Wiktionary and ReVerb system, which is an open information extraction tool that extracts binary relationships of type *phrase-relation-phrase* in an unsupervised manner. The whole database contains 414 thousand English concepts and 903 thousand relationships between them.
- **Semantic Network of Terms** (SemNet) [2] is a large-scale network of technical terminology which allows querying terms and retrieving ranked lists of their semantically related terms. The network was automatically constructed based on the noun terms from English Google Books

Ngram Dataset using word co-occurrence analysis. The network consists of 2.8 million distinct single and multi-word terms and 37.5 million weighted edges between them. SemNet includes a large part of the same concepts and relationships from similar semantic knowledge bases such as WordNet [3] and ConceptNet [1].

- **TF-IDF**, term frequency-inverse document frequency is a general numerical statistic that defines the importance of each word in a document collection. It is often used in information retrieval and text mining as it gives good string-based performance.
- **Categories** Each post contains associated categories. In our dataset there are three hierarchical category types. Their values are more thoroughly explained under the dataset description (Section VI-A).

VI. Evaluation

TODOwrite section intro

A. Dataset

We reviewed many question answering (QA) systems and datasets. Finally, we decided to use Yahoo! Answers Webscope L6 dataset [4] because it contains a broad range of distinct question categories and there are enough active users in each of the categories we selected. Other acceptable datasets could be extracted from sites such as AskVille (http://askville.amazon.com), Mahalo (http://mahalo.com), Quora (http://quora.com) and StackOverflow (http://stackoverflow.com). The latter also includes a lot of users but it does not have specified specific categories and more importantly, the whole dataset is mostly single domain oriented. Furthermore, there are lots of other systems like AllExperts (http://allexperts.com), Ask.com (http://ask.com), Answers.com (http://answers.com) and others that lack one or more public features to build useful dataset.

Yahoo! Answers is a QA site where people post questions and answers, all of which is publicly available to any web user. The L6 dataset was collected from this site in 2007 and it includes all the questions (i.e. 4483032) and their corresponding answers. Next to these data some anonymized metadata is included, so that we can model concepts of post and user.

Instances of post type represent questions and answers. Each instance consists of text in the body, selection of main category, category and subcategory and the id of the user who wrote the post. Questions additionally contain also title. For answer posts, the owner user is known only if it was considered the best answer for the question.

User instances are completely anonymized, so therefore we model them using their id and we also count how many answers or questions they posted in specific category type.

From the full dataset we extracted three database types:

- **Type 1**: This dataset models interest as it contains users that asked at least ten questions and each of these questions must have at least five answers.
- **Type 2**: To model knowledge, we extracted users who best answered at least ten questions. Again, each question thread needed to have at least five answers.
- **Type 3**: To jointly represent interest and knowledge, we extracted users that asked at least five questions and best answered at least five questions. Also, each of these threads needed to have at least five answers.

Each dataset type contains 100 users which are distinct across the datasets. As there are many different categories in the data, we representatively selected them, so that extracted 100 users within each dataset are formed as shown in Table I.

Table I. Distribution of

Category	Number of selected users
Society & Culture	35
Food & Drink	35
Computers & Internet	15
Travel	10
Cars & Transportation	5
Total	100

TODOspecifics in code, evaluation setting

B. Results

TODO

To evaluate our work we use metrics mean reciprocal rank (MRR) and Precision@N (P@N) which are widely used in the field. They are defined as follows: (1) MRR is a measure to evaluate information retrieval task in which a list of possible responses to a query is ordered by a probability of correctness. The score is defined as the average of the reciprocal ranks for a set of queries Q,

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}.$$

(2) Precision is the fraction of the documents retrieved that are relevant to a query. P@N is therefore precision which is evaluated at a given cut-off rank N (i.e. considering only top N results). In our domain it measures the proportion of correctly selected experts among all experts and a probability of how likely asking one of the top N will result in getting a correct answer.

Visualization: - ascending graphs for P@N - MRR table: x=db type, y=evidence types used

C. Discussion

VII. Conclusion

The conclusion goes here.

Acknowledgment

The work has been supported by the Slovene Research Agency ARRS within the research program P2-0359 and part financed by the European Union, European Social Fund.

References

[1] R. Speer and C. Havasi, “Representing general relational knowledge in conceptnet 5.” in *Proceedings of Language Resources and Evaluation Conference*, 2012, pp. 3679–3686.

[2] H. Agt and R.-D. Kutsche, “Automated construction of a large semantic network of related terms for domain-specific modeling,” in *Advanced Information Systems Engineering*, C. Salinesi, M. Norrie, and O. Pastor, Eds., vol. 7908. Springer Berlin Heidelberg, 2013, pp. 610–625.

[3] G. A. Miller, “Wordnet: A lexical database for english,” *Communications of the ACM*, vol. 38, pp. 39–41, 1995.

[4] M. Surdeanu, M. Ciaramita, and H. Zaragoza, “Learning to rank answers on large online qa collections,” in *In Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies*, 2008, pp. 719–727.