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 auestion 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

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

V. UserProfiling

A. Sources

TODOwrite about different sources. There are three types of sources: System (categories), Semantic (...), String (TFIDF). They are divided if extracted from Title or Body and can refer to Question, Answer or the whole thread.

- Categories (Cat1, Cat2, Cat3)
- ConceptExtraction
- TFIDF
- SemNet

VI. Evaluation

A. Dataset

-say something about other datasets?

DATASET: The dataset contains 4483032 questions and their answers. All the data is anonymized with some additional metadata. For each question we have: anonymized question URI (e.g. 432470) subject optional content best answer all n-best answers question categorization using taxonomy ManCat-¿Cat-¿SubCat (e.g. Education & Reference -¿ Trivia -¿ Trivia) ID of person asking the question ID of user providing best answer optional language of posted question optional location of posted question date and timestamp of the question The data is from 2007. The L20 dataset contains data from 2006. Are the users anonymized in the same way? *Note: The scientific paper using Yahoo! data should mention phrase "Yahoo! Webscope" and when the paper is

published, a copy of the paper should be send to research-data-requests@yahoo-inc.com. Source: Yahoo! Webscope

OUR DATASET: question,answer = we need the whole threads.

1. Users that only asked - the current db: TYPE 1 DB topQuestioners (interest) - the one who ASKED 10(questions)+1(answer) questions(it was 5-20). Each of these questions must have at least 5 to max 20 answers. (from these result I only used a part of it -¿ 200 questioneers). Shrink this dataset to 100 users * 10 questions = 1000 questions. For that 100 users, find also all answers they have answered (the whole threads).

There are 4000 questions and 3607 best answerers. (some of them - small numb - were best answerers for more than 1 question - less than 20 users)

For experiment - for theese users find questions where they BEST ANSWERED. So we will evaluate them using those questions.

2. Users that only answered - TYPE 2 DB topAnswerers (knowledge) - the one who BEST ANSWERED 10(answers)+1(answer) questions. Each of these questions must have at least 5 to max 20 answers. Again we need about 100 best answerers within 1000 questions.

3. Users that both ansked and answered - TYPE 3 DB topAskersAndAnswerers (knowledge&interest) - the one who ASKED and BEST ANSWERED 10(5*questions+5*answers)+1(asnwer). Each of these questions must have at least 5 to max 20 answers. Again we need About 100 users within 1000 questions.

For all DB types - 1 additional question where they best Answered - for Evaluation.

I think that it is the best to stick to some constant (e.g. as I wrote 10) for the number of questions (that asked or answered) by which we select users.

Also all of these sets - topQuestioners, topAnswerers, and topAskersAndAnswerers - should not have intesection (different users). Because we will mix them for experiments. Put all DateSets in the same DB, just assign different atribute - topQuestioners, topAnswerers, and topAskersAndAnswerers Profile all with Evidences: TFIDF, ConceptExtraction, SemNet, Categories

Rezultati extractanih podatkov: Number of examples by Category (for EACH TYPE): ”Society & Culture”: 35 ”Food & Drink”: 35 ”Computers & Internet”: 15 ”Travel”: 10 ”Cars & Transportation”: 5

SELECT COUNT(*) FROM interestminingl6typeall.Posts WHERE postTypeId = 1; #3275 SELECT COUNT(*) FROM interestminingl6typeall.Posts WHERE postTypeId = 2; #38034 SELECT COUNT(*) FROM interestminingl6typeall.Posts WHERE postTypeId = 2 AND ownerUserId != ”; #3269 SELECT COUNT(*) FROM Users; #300

B. Results

In the expert searching area, the metrics of mean of average precision (MAP) and Precision@N (P@N) are widely used to evaluate the precision of expert searching. They are defined as follows:

DEFINITION 1. MAP: MAP is the mean of the average of precisions over a set of query questions. The average of precisions for a query is the average of precision at each correctly retrieved expert.

DEFINITION 2. P@N is the percentage of top N candidate answers (experts) who are correct. (N=1,5,10) Precision measures the proportion of correctly selected experts among all experts (N).

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] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.