Posudek vedoucího diplomové práce

Jméno a příjmení autora posudku:	RNDr. Pavel Pecina Ph.D.
Jméno a příjmení autora práce:	Feraena Bibyna
Název práce	Query expansion for medical information retrieval

Text posudku

The work presented in the thesis belongs to the area of information retrieval and focuses on the methods for query expansion and similar techniques applied to the medical domain.

The work was motivated by several factors: First, searching for medical information becomes very common in the current Internet era when more and more people seek for health-related advices on-line before (or instead of) going to see a doctor. Second, the amounts of medical information available on-line is increasing rapidly; its quality varies and delivering relevant and reliable (and appropriate) information is difficult. Third, medicine as a long-studied field can provide large and high-quality terminological resources (thesauri and semantic networks) which can be used to improve search quality in this domain. Feraena Bibyna in her thesis investigated several methods based on reformulation and modification of users' queries, such as query expansion, blind relevance feedback, and query term weighting. She experimented with various semantic relations and also explored the effect of interpolating outputs of several retrieval models.

The thesis is structured into seven sections including introduction and conclusion. After the introduction, the author presents some theoretical foundations related to the work. Then she describes the data set used in the experiments and continues with description of the methods and models employed in the work. This is followed by sections presenting and discussing the experimental results on training data and two sets of test data.

The experiments were conducted on the document collection and queries provided for the CLEF eHealth shared task in 2014 and 2015. The queries in the two years of the shared task differ to a large extent (they model two different use cases) and so the two sets of queries were joined and then randomly split into one training set containing queries from both 2014 and 2015 and two test sets with queries from 2014 and 2015 separately. Thus, the results are not directly comparable to the official results of the shared task but this step reduced the problem of having too diverse sets of queries. The retrieval systems were trained on the mixture of the two query types but tested independently for performance in the two use cases.

The text of the thesis is well written and readable, the methods well described and illustrated using appropriate examples. The amount of experiments performed is exceptional and the author managed to clearly present their results. The author independently and thoroughly studied large number of methods and proposed several modifications which showed some interesting improvements of search quality, although they were not confirmed on both test sets. The results are well discussed and analysed. The thesis builds on the particiation of the author in the CLEF eHealth shared task focused one medical information retrieval. The team placed third among 12 teams in total, which is a very good result and a good starting point for future work. I recommend the thesis to be defended.

Doporučení k obhajobě

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Soutěž studentských prací

Vynikající práce vhodná soutěže studentských prací: **NE**.

V Praze dne 31. 8. 2015

Podpis: