Ph.D. Thesis Opponent's Review

Thesis title: Mathematical Search Engine

Author: Jozef Mišutka

Reviewer: Jiří Dvorský

Thesis Structure

The thesis is conceptually divided into six chapters. After introduction in chapter 1 a brief overview of general theory of search engines is given in the chapter 2. Chapter 3 provides excellent state of the art and latest development in the area of mathematical search engines. Chapter 4 describes the main author's contribution in the thesis – design and implementation of EgoMath search engine. The next chapter, chapter 5, contains author's extenstion of the feature based mathematical search engine.

The thesis brings comprehensive study of present knowledge of searching of mathematical formulæ on the web. The thesis also presents result obtained using EgoMath search engine which was succesfully tested on large data. The presented complex comparison of two mathematical search engine is probably the first one in the scientific literature.

I think that research goals of the work specified in introduction are fullfilled. Author presents new ideas or extends ideas of another scientists.

Comments and questions

I have a few questions:

1. Page 46, top. Succes of your search engine measured by number of returning visitors may be measure of commercial success, not scientific measure.
2. Equation (4.6.1) on page 66. There are a lot of coefficients, such as $CF(q, d)$, but there are no serious explanation how they are defined or computed from the data. Please define them.
3. Section 4.9 Evaluation. In the whole section there is no data about classical Precision/Recall measure of information retrieval system. How do you know that formulæ retrieved by your system is relevant? There are no false hits? And all relevant formulæ were retrieved. Data about size of the index and time to build the index are interesting, but P/R data will be helpfull.
4. Page 93, bottom. What is the interpretation of this formula?

$$1 - \frac{2\arccos(similarity)}{\pi}$$

or

$$1 - \frac{2}{\cos(similarity)\pi}$$

5. Section 5.3. Comparison of the EgoMath and FBA. There is the same problem as in the evaluation of EgoMath. These two search engines provides mostly the same results. But how do you know that these results are correct? If both search engines provide the same incorrect results, comparing their outputs will turn out well.

Author's Publications

The author does not explicitly provide list of his publications in the thesis, but they may be easily found in WoS or Scopus databases. In WoS and Scopus there are 3 and 2 publications respectively. I think that these numbers and quality of publications are still acceptable for Ph.D. student, but it is minimal level of publication activity. But I have one question: why is there such a gap between years 2008 and 2011 and from 2011 till now? There is no publication for these ? No research? I think that thesis based on two or three publications (well, with exception of two papers in Nature journal) is not good idea.

Formal Aspects of the Thesis

Formally, the thesis is carefully prepared, using very good and clear English. Tables and especially figures with graphs are clear and esy to read. Typography of the thesis is on high level, with the exception of a several widows e.g. on pages 10, 28, 33, 64, and 64.

Conclusion

I am convinced that the presented Ph.D. thesis represents mature study providing valuable contribution to the state of scientific knowledge in the area. I recommend the thesis for defense.

Olomouc, August 21st 2013

Jiří Dvorský
Department of Computer Science
VŠB – Technical University of Ostrava