Annex 6: Habilitation thesis reader's report

Masaryk University
Faculty   MU Faculty of Informatics
Field of Habilitation   Informatics

Applicant   RNDr. Vlastislav Dohnal, Ph.D.
Affiliation   Masaryk University, Faculty of Informatics
Habilitation Thesis   Developing Similarity Search Technology

Reader   Professor Gonzalo Navarro
  Affiliation   Faculty of Physical and Mathematical Sciences, University of Chile

Report Text (as large as the reader deems necessary)

The applicant has a quite good volume of scientific production, including some top-tier conferences like ICDE and journals like MTAP. An indication of impact is his h-number of 8 in Google Scholar, which is reasonable. It is probably more remarkable that his most highly cited article (that in MTAP) has received almost 100 citations. Another indication of impact that is not well captured by this measure is his coauthorship in the book "Similarity Search: The Metric Space Approach", one of the two main books in the area of similarity search. Together with this extremely important publication, there are several other book chapters and survey articles, at least three of which appeared in quite influential venues (in the Encyclopedia of Database Systems and in SIGSPATIAL Special).

The applicant has managed to explore a wide set of problems related to similarity search, and his name has become well known in relation to good solutions to various problems. His MTAP article (100 citations), for example, is about the D-index, a well known indexing structure for metric spaces. His second most highly cited publication, in DEXA 2003, was about a technique to carry out metric joins, a problem I believe the applicant was a pioneer in considering in the area of metric spaces. His third article, that in ICDE 2008 (an extremely competitive conference), is about a novel and very intriguing topic: using metric tools to explore social networks. Once again, I believe the applicant is the first in considering such an idea. The fifth is about MUFIN, a real similarity search system. The applicant has been successful in following a path where good algorithmic ideas are not decoupled from practical implementations, and this is very valuable. Another relevant article, in ISM 2008, addresses the problem of similarity searching over P2P networks. Considering massive data sets and practical tools to handle them is characteristic of the applicant's work.

I have done this short tour over his most influential papers to illustrate how the applicant has succeeded in carrying out good-quality research in several topics of his interest: basic metric data structures, metric databases (e.g., metric joins), massive metric datasets (e.g., parallelism, distributed solutions, secondary memory structures), real applications (e.g., MUFIN and various image database products that actually work), social networks, and so on.

I think these achievements are more than sufficient to meet the requirements for a habilitation thesis in Computer Science. My recommendation for the applicant, considering what is next in his career, is to try to develop further his independence. There are some coauthors that

appear almost always in his publications, especially his former advisor, Prof. Zezula. It is very nice to keep cooperating with the former advisor (this is what we, advisors, dream about!) but it is important that the applicant *also* shows that he can follow independent research projects. As said, this is not yet a problem for the habilitation, but will become relevant for his next promotions.

Reader's questions to answer to defend the habilitation thesis (number of questions is upon reader's consideration)

1. Recently, in a SISAP invited paper, Prof. Tomas Skopal wrote a provoking article questioning the usefulness of the metric space approach, at least if used in naïve form, in practical cases. I presume you know the article. Could you summarize it, and add your personal viewpoint on this topic, in particular in connection with your own work in building actual similarity retrieval systems?

2. Can you give your perspective on how you see the field, where is it going, what are going to be the relevant problems in the next 10 years? Will they still be basic algorithmic problems as in the past decade, or more than that, problems of managing scalability, or retrieval quality, or others? I am thinking on the path that research on textual information retrieval has followed in the last 20 years, for example, and what are the important problems of the search engines today. For text it seems that (at least basic) searching is not anymore an algorithmic problem, but maybe similarity search is not so mature?

Conclusion

Vlastislav Dohnal's habilitation thesis of "Developing Similarity Search Technology" *does* meet the standard requirements for a habilitation thesis in the field of Informatics.

In Chile on Jan 16, 2012   Gonzalo Navarro