Context-based Improvement of Search Results in Programming Domain

Jakub KŘÍŽ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
jacob.kriz@gmail.com

When programming the programmer faces a great deal of many different problems related to his work. He often uses web search engines in order to try and solve them.

When searching the web the users usually enter only a few words to form a search query, which may often lead to unsatisfactory results [1]. Programmers are expected to be usually better at putting together a search query, however, because they make the searches much more often than regular users it is likely that they get inattentive and make searches with unsatisfactory results as well.

Programmers do many things during a typical programming session apart from just writing the source code. They do all these things with a certain intention, in a certain context. By understanding this context we can make the results of the programmers' web searches more accurate and relevant. In this work we propose a method for extracting context metadata in the programming domain, from the sources specific to programming, in order to build a programmer's context model.

Based on our experience with the programming domain we group the problems programmers usually face into three categories:

- *Conceptual problem* – the programmer is trying to understand the idea of an algorithm or come up with an algorithmic solution to a problem
- *Technical problem* – the programmer is solving a problem associated with using the methods of the language or library or API methods which he is currently using
- *Error* – the programmer is solving an error which occurred

The context model should represent the programmer's intentions in any given point in time. We base the structure of this model on the types of problems we described:

- Conceptual part
- Technical part
  - Language

* Supervisor: Tomáš Kramár, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 69–70.

  - Framework / libraries
  - Key identifiers
- Error part
- Programmer's state

The *conceptual part* says about the conceptual intentions of the programmer. It is composed of a set of ranked keywords. These keywords are extracted from the active part of the source code, the part with which the programmer currently works, using a custom tf-idf algorithm, which is modified to extract conceptual keywords and only from the active part of the code. Similar extraction method for conceptual keywords has been successfully used in previous works [2].

The *technical part* includes information about the technologies the programmer is currently working with. The *language* part contains the identifier of the programming language; the *framework/libraries* part contains the identifiers of currently used frameworks and libraries. The *key identifiers* part is composed of a set of ranked key identifiers of core, library or API methods and classes. They are extracted using a similar, modified tf-idf algorithm as the conceptual keywords. The *error part* contains the identifier of the last error which occurred.

The *programmer's state* part says in which state the programmer is currently in. The state is detected by analysing the programmer's activity in the IDE. The method considers several factors, like the time of his last typing or time since last compilation and uses supervised learning algorithm to classify his state.

We further use the context model to improve the results of programmer's web searches. We do this by reranking the search results. Based on the detected programmer's state the method picks the relevant part of the context model. When ranking the search results the method boosts the rating for documents, which contain words or identifiers which were also found in the context model. We believe the context model, as proposed is very versatile and can possibly be used in many ways.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 492-497.

Acknowledgement. This contribution is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

[1] Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 2000, vol. 36, pp. 207–227

[2] Ohba, M., Gondow, K.: Toward mining "concept keywords" from identifiers in large software projects. In: *Proceedings of the 2005 international workshop on Mining software repositories.* MSR '05, New York, NY, USA, ACM, 2005, pp. 1–5.