Online Gathering of Information from Text Sources

Štefan SABO*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
xsabos@is.stuba.sk

The project I am working on is dealing with gathering of information from text sources available on the web. The information wouldn't be gathered from databases, or already processed sources, but from random sources presented in natural language. These sources may come in many forms, like consumer forums, articles, user comments or various blogs. The aim of this project would be to create a system, able of locating, harvesting and subsequently analyzing such sources. This analysis would be able to provide us with information about various products, events, etc. The first idea would be to implement opinion mining about consumer products as cell phones, notebooks, or other electronics. Although this is just a first notion and could be subject to a change further down the road.

The rough schema of the system is presented on Fig.1. The system receives queries from users, processes them and produces results, which are returned back to the user. The data is gathered by the means of web crawling bots and stored in a database. After being gathered, the data is analyzed and the results are produced. The individual steps presented in this diagram don't need to be performed in this order. The returning of a result will always be preceded by issuing a query. However, the system-internet and the system-database interaction need not occur only after issuing a query. If we can specify, what general kind of data needs to be gathered before the query is issued, the data may be gathered on-the-fly all the time. Also certain parts of analysis can be done independently, without the knowledge of the query. Examples of such steps would be the author analysis, parsing, or syntactic preprocessing.

Similar systems dealing with opinion mining are no new idea [1, 2]. However, this system would differ in the means of information locating and gathering. To achieve reasonable reliability we need to analyze a rather wide database. In order to do so, I would like to utilize the social insect model, namely the metaphor of bees gathering food for their hive. The bees would be implemented as web-crawlers.

* Supervisor: Anna Bou Ezzeddine, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 67–68.

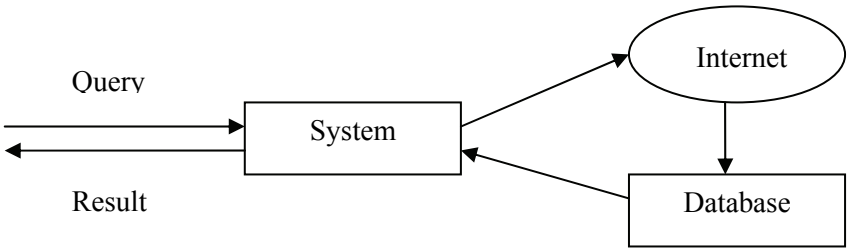

Fig. 1: Schema of the system interaction

These crawlers would be able to search the web and download pieces of information for further processing. The advantages of this approach consist in the ability of web crawlers to make decisions and prefer the more suitable sources over the ones that are less interesting. The new information tends to be linked to more often than the older one. So the web crawlers searching the web are more likely to localize such a source, just like a bee stumbles upon a quality food source. These sources would subsequently be harvested, meaning that the text information would be downloaded to a database and ready to be analyzed.

The analysis would provide us with opinions about chosen products. After the implementation, series of tests would be performed. Aim of these tests would be to compare different strategies, or parameters and find out, how good would be produced.

Acknowledgements. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010

References

[1] Spangler, S. et.al., COBRA - mining web for COrporate Brand and Reputation Analysis, Web Intelligence and Agent Systems,Vol. 7, IOS Press, Amsterdam, 243-254, 2009.

[2] Cai, R. et.al, iRobot: an intelligent crawler for web forums, Proceeding of the 17th international conference on World Wide Web, ACM, New York, 447-456, 2008.