Intelligent Information Retrieval Tools for Police

Nishant Kumar[1], Jan De Beer[2], Jan Vanthienen[1], and Marie-Francine Moens[2]

[1] Research Center for Management Informatics,
Katholieke Universiteit Leuven, Belgium
{nishant.kumar, jan.vanthienen}@econ.kuleuven.be
[2] Legal Informatics and Information Retrieval group,
Katholieke Universiteit Leuven, Belgium
{jan.debeer, marie-france.moens}@law.kuleuven.be

Abstract. Intelligent information retrieval tools can help intelligence and security agencies to retrieve and exploit relevant information from unstructured information sources and give them insight into the criminal behavior and networks, in order to fight crime more efficiently and effectively. This article aims at analysing off-the-shelf information extraction tools on their applicability and competency for such applications.

1 Introduction

With increasing volume of crime data, intelligence and security agencies across the world need intelligent support systems which can help them to retrieve and exploit relevant information and give them insight into the criminal behavior and networks, in order to fight crime more efficiently and effectively. The unstructured information (e-mails, reports, web pages, etc.), representing the bulk of all information, poses a great challenge in automation.

Many of the IR tools available today provide good and fast solution to the retrieval problems (retrieval, querying, structuring, visualization, extraction, etc.). But it is also very difficult to know which of these tools are most effective for a given application. We report our work on the evaluation of 10 tools, shortlisted from a market selection of 23 tools, under the INFO-NS [1] project for the Belgian Federal Police.

2 Evaluation Method

We identified several user profiles, their functional requirements and priorities and generalized them over user profiles to five high-level *use cases*, namely (free text search, Metadata Search, Classification, Named Entity Extraction and Entity Linking). For each of these use cases, we compiled a detailed evaluation form based on sound evaluation frameworks, covering three crucial aspects of assessment, Conformity[1], Quality, and Technical.

We tested the use cases on multilingual document collection containing more than half a million real-life case reports, in Dutch and French, encoded in the MS Word file format.

[1] Visit AGORA at http://www.belspo.be/belspo/fedra/prog.asp?l=en&COD=AG

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 664–665, 2006.

3 Evaluation Results

Free Text Search: The proprietary fuzzy matching algorithm of one tool gave excellent results on most of the variation types considered, whereas the use of the Soundex ([2]) and edit distance ([3]) operators as provided by most other tools proved to be ill-suited for most variation types. Moreover, neither Soundex nor edit distance copes well with word reorderings, e.g. as with person names.

On relevance ranking ([4]) one tool consistently produced well-ranked result lists (on a scale from 0 to 100, baseline+30 up to +70) and whereas other tools clearly showed variable scores.

Metadata Search: Standard document attributes (such as url, title, author, date, size), when available, are automatically imported by the tools. Most information retrieval tools also derive a static summary simply by extracting the most salient sentences or phrases from the text.

Named Entity Extraction: IR tools support NE extraction ([5]) on common entity types person names, organisations, locations, and time instances. Results show high precision, up to 97%, on the most common entity types, while recall is very poor, less than 50%.

4 Conclusions and Future Directions

This evaluation of tools has given us the impression that without a careful consideration of functional requirements and integrating a good human computer interaction feature these tools might prove to be of less fruitful. We have also found that the tools lack support for cross lingual search, an important aspect in a trilingual country like Belgium. While in case of entity extraction and classification most of the tools work on keyword matching, which does not give the right contextual result and therefore decreases the relevancy factor and also they lack support for noisy texts.

References

[1] N. Kumar, J. De Beer, J. Vanthienen, and M.-F. Moens, "Multi-criteria evaluation of information retrieval tools," in *Proceedings of the 8th International Conference on Enterprise Information Systems(ICEIS)*, 2006.
[2] R. Russell and M. Odell, "Soundex," Patent 01 261 167, 1918.
[3] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, no. 4, pp. 845–848, 1965.
[4] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the ACM SIGIR Annual International Conference on Information Retrieval*, vol. 27, July 2004.
[5] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," in *Proceedings of the International Conference on Intelligence Analysis*, 2005.