Assessing the State of the Art of Commercial Tools for Unstructured Information Exploitation

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

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

Abstract. This paper provides a snapshot of the state-of-the-art in information retrieval and information extraction from text through a selection of commercial, market-leading tools. Rooted in a research project conducted for the Belgian Police, we give an overview of the main (desired) features provided or lacked by these tools, along with their measured quality in operation. Various shortcomings and suggestions for improvement will be formulated.

1 Introduction

Since it has become so easy to create, share and store information in today's pc-centric culture, the rate of information is now growing exponentially, to the extent that our ability to create information has substantially outpaced our ability to retrieve and exploit relevant information. Amongst many others, this fact challenges intelligence and security agencies, which strive on discovering and dispersing the right information to the right people well in time. In addition, they aim at discerning patterns that give valuable insights into novel criminal behaviour and organisation, so as to help fight crime more efficiently and effectively.

The problem is more pronounced for unstructured information, representing the bulk of all information, examples of which are text (e-mails, reports, web pages, etc.) and other multimedia documents (sound, pictures, graphs, video, etc.).[3] This is because the information is embedded in a language or format that is more easily processed and interpreted by the human brain, and generally difficult to process by automated means. This is opposed to structured information (database records, spreadsheets, metadata, etc.), that comes in rigid, well-defined forms that are readily exploitable by computer systems. In this paper, we will restrict ourselves to the exploitation of textual, unstructured information.

[3] According to a Delphi Group white paper ([1]), the percentage of unstructured data to the total amount of data is estimated at 85% and is growing.

The challenges posed fuel academic and corporate endeavour at developing suitable tools for information exploitation. However, in general these tools are tremendously expensive. In addition, whilst being labelled with promising marketing statements on their usefulness and abilities, their vendors are generally reluctant to provide the software free of cost for evaluation purposes. For these and other reasons, great value may be attached to (comparative) studies conducted by independent parties with sufficient domain knowledge to assess the true quality, performance, and abilities of these tools. Through our research in the INFO-NS project for the Belgian Police (BP) and this paper in particular, we contribute to this line of work. Our findings have direct relevance to both end users, decision makers, and software providers, in that objective information and realistic expectations can be acquired, and that formulated suggestions for improvement may bring the tools closer to end users' desires and needs.

The rest of this paper is organised as follows. In the next section we elaborate on our selection of evaluated tools and the methodology used in assessing them. We then give an overview of our main findings with regard to the selected tools on a number of evaluated use cases. We end with our conclusions and mention related work for further reading.

2 Evaluation Methodology

2.1 Requirements Analysis

In order to reveal the profile of the exploitation tool desired and suitable for use within the BP, we initiated our project with a market study paired with a thorough site study ([2]). Through this survey, we identified the different information sources, user profiles, functional and technical requirements. On a high level, the functional requirements can be grouped in the following *use cases*.

Free-Text Search. Search queries defined in some query language are matched against the textual contents of a document collection in order to retrieve and rank the most relevant documents (cf. web search engines such as Google, retrieving relevance-ranked web pages).

Metadata Search. Structured data that is associated to documents is turned into additional, searchable document fields (also called attributes, e.g. title, author, date), next to the free-text document contents.

Classification. Documents are automatically classified into either a prebuild or an automatically derived taxonomy (a topic hierarchy) based on content analysis and similarity.

Named Entity Extraction. Mentions of named entities are automatically recognised in text, disambiguated and classified into any of the supported entity types (e.g. persons, organisations, vehicle numbers, time).

Entity Linking. Extracted entities in a document set are linked and visualised in networks (e.g. co-occurence graphs), which serve to analyse the relationships between the entities.

2.2 Selection of Tools

With sufficient insights into the requirements and a global overview of the market, we made an initial selection of 23 reputable commercial tools.[4] We were able to further reduce this list through product information coming from various sources, including brochures, company web sites, similar studies (confidential), and meetings with company representatives and technical engineers. Ultimately, 10 tools were selected for active participation.[5]

On a functional level, we can assign (non-exclusively) any of the selected tools to three tool categories. *Information (document) retrieval tools*' primary purpose is to allow exploring and retrieving information (documents) from any of the information repositories managed by the tools (i.e. document collections or the World Wide Web). *Data mining tools* offer an array of data input, output, and processing modules that can be combined in a single scheme for the processing of structured, tabular data. The selected data mining tools each provide a text mining module, which produces a tabular index from selected text documents, listing word (groups) along with frequency counts and POS tags.[6] Lastly, *information extraction tools* semantically classify information that is literally contained in text, such as entities (persons, companies, products, quantities,...), facts (properties of entities or entities in relationship), actions, scenarios, etc.

2.3 Evaluation of Selected Tools

For each of the use cases, we compiled a detailed evaluation form along with semi-automated evaluation procedures emerging from self-developed evaluation models ([2, 3]), covering three crucial aspects of assessment.

Conformity evaluation measures to which extent functional and technical requirements are met by the tools. As we found out that functional support is hardly a yes/no answer, exceptions, side conditions, alternative options, and remarks were noted down.

Qualitative evaluation measures the quality (such as usefulness, relevance, or accuracy) of the results as output by the tools on a number of carefully drafted test cases.

Technical evaluation measures the allocation of system resources (such as processor, memory, diskspace, network bandwidth, runtime) by the tools on a representative set of (scalable) test cases. In addition, ease of installation and configuration, indexing statistics and errors, robustness and stability were noted down.

[4] We did not concern ourselves with the state-of-the-art of academic nor freeware tools. By nature, their documentation and evaluation is much more open to the public.

[5] By contractual obligations, we cannot publish any names or details that might reveal the identity of the evaluated tools in our publications.

[6] Part-of-speech (POS); one of the traditional categories of words intended to reflect their functions in a grammatical context, such as verbs, adjectives, prepositions, etc.

Amongst the materials provided by the BP for the test cases is a multilingual document collection containing half a million documents of real-life case reports, Dutch and French, encoded in the Microsoft Word file format. This collection constitutes a representative sample of the full operational collection, which is expected to hold a number of documents that runs in the order of ten million.

A typical case report consists of a header holding information fields, minimally including a unique, systematic identifier termed 'PV number', the originating police zone and contact information. The header ends with a formal clause specifying the police officer(s) and date of writing. Multiple sections follow, in which a verbal description for the type(s) of crime committed and the detailed identity of all involved persons (victims, suspects, witnesses,...), cars, objects,... are listed in template-style. More sections follow, containing brief and detailed free-text descriptions of the facts, along with literal transcripts of hearings in the subject's preferred language, and sometimes followed by evidential materials such as pictures (very rare in our collection), official documents, and so on. Each case report concludes with a formal closing statement. The free-text sections are typically noisy, containing spelling errors, partial, faulty or phonetic entity names, inconsistent abbreviations, capitalisation, spacing, interpunction, accented characters, etc. As a result of BP's authoring tool for case reports, one case report is typically dispersed over multiple physical text files, which we refer to as *documents*. For example, the main content is separated from every appendix, giving rise to a series of consecutively numbered documents.

All test cases were performed on a single computer, hosting both the server and client components of each tool. The machine is equipped with an Intel dual processor (2×1Ghz), 2GB (gigabytes) of volatile memory, 4GB of virtual memory, a SCSI RAID-0 disc array with separate partitions for the operating system Microsoft Windows 2003 Server (supported by all evaluated tools), the program files, and the data.

3 Evaluation Results

For each use case, we mention the main application purpose(s) in view of the BP's requirements, the common and differing support that is provided by the evaluated tools, general findings on their qualitative assessment, followed by a number of attention points (criteria) that were found to be determinant in the comparison of tools, and finally some suggestions for tool improvement.

3.1 Free-Text Search

Purpose. The basic objective of the BP in the INFO-NS project is to find a good document retrieval system to search and access the central repository of documents for operational use (i.e. case reports). This way, documents are not just accessible through querying the corresponding structured data, but also become accessible through open-ended search queries posed against the actual, complete document contents.

Support. All tools offering document retrieval have a comparable client-server system setup. They support all major document file formats (including TXT, HTML, PDF, productivity packages like Microsoft Office) and repositories (including file systems, databases, document management systems like Lotus Notes, and the World Wide Web). When crawled, a document's descriptors (metadata, words, noun groups, classification codes, etc.) are included or updated in the tool's proprietary index structures, a process referred to as *indexing*.[7] Most tools now allow for flexible deployments; workload, content, and indices can be distributed over multiple servers to keep the entire system scalable and performant.

Searching for documents works very similar too. Each tool defines a query language in which the user expresses her information needs. The language minimally supports the specification of keywords and phrases, implicitely or explicitely coined using search operators and/or modifiers (boolean, proximity, wildcard, fuzzy, lemmatisation, and thesaurus operators, case sensitivity and language modifiers, etc.). In addition, conditions on metadata can be posed, and the search scope bounded by selecting any of the target document collections that were previously indexed. Each tool employs its proprietary relevance ranking model that scores each target document in accordance with the search query. The results are returned to the user and presented in (paged) document lists or navigable classification (metadata) folders, holding links to the actual documents.

No evaluated tool supported crosslingual search out-of-the-box. With crosslingual search, queries posed in one language retrieve relevant documents regardless of their content language. Integration of third-party machine translation software or crosslingual dictionaries paired with existing query expansion mechanisms may offer solutions. However, both translation software and dictionaries are currently still limited in quality and completeness.

Quality. Text extraction and markup removal was found to work fine for some of the most common document types, as tested through the literal search query type (supported by all tools). However, some tools discarded Microsoft Word page headers and footers for indexing.

In line with the BP's requirements, emphasis was given on fuzzy search capabilities, in order to recognise common variations (or noise) with regard to terms and entity names that are commonly found in case reports (cf. supra). The proprietary fuzzy matching algorithm of one tool was found to give excellent results on most of the variation types considered, whereas the use of the Soundex ([4]), the edit distance ([5]), and wildcard operators as provided by most other tools proved to be ill-suited for most variation types. Soundex allows for some reasonable character substitution errors, but is clearly targetted at English phonetics, and does not tolerate consonant additions or removals (as with abbreviations). Edit distance considers all three types of character errors (removal, insertion,

[7] Depending on the tool, indexing 1GB of Microsoft Word documents took 43 minutes up to 2 hours, requiring 40, 50, 80 or more percent of index space, with no clear observed correlation on search quality or speed.

and substitution), but does so in a very uniform way, assuming all character errors are independent from one another. In practice this requires large thresholds in order to recover abbreviations, interposed or omitted words, interpunction,... whereas small thresholds are desired to limit spurious matches as much as possible. The effective use of wildcard operators - single and multi character sequence - is limited to special cases only, as they constrain, respectively widen the search in a scarcely controlled manner. Lastly, none of the above operators copes well with word reorderings, e.g. as with person names.

The relevance ranking model was evaluated using the *rpref* metric ([3]); a generalisation of the *bpref* metric ([6]) towards graded relevance assessments.[8] In short, the metric measures the extent to which less relevant documents are ranked before more relevant documents. Compared to a baseline *rpref* score of a random ranking model, one tool was found to consistently produce well-ranked result lists (on a scale from 0 to 100, baseline+30 up to +70) and promote a priori determined highly relevant documents, whereas other tools clearly showed variable, and some language-dependent *rpref* scores (below versus above baseline) and recall of the a priori determined documents.

Criteria. Tools differ most when it comes to coverage, depth, and quality of standard included resources (language packs, synonym lists, abbreviation dictionaries, thesauri, etc.).[9] Another crucial factor is the expressivity of the query language (extending from simple keyword search), carefully balanced with ease of use, hiding query language syntax and complexities underneath a clear and intuitive search interface. Lastly, the quality of results in terms of precision, recall, and ranking stems from the implemented relevance ranking model.

Suggestions. We emphasise the need for build-in support of crosslingual search, as many organisations nowadays have to deal with this problem. For the BP, tools should also be capable of handling documents containing text in different languages (hearings e.g.). Furthermore, we encourage research into and the implementation of multipurpose fuzzy search operators with user-definable threshold value, other than the rarely suitable, yet popular Soundex and edit distance operators.

On the retrieval side, we encourage experimentation with alternative retrieval models, different from the keyword-based retrieval models that are implemented by all evaluated tools. Various interesting retrieval models have been proposed in the academic world, but they do not seem to find their way outside the laboratory. As one instance, the XML retrieval model ([7]) could be used to exploit document structure in finding relevant sections. Structure is inherent in most documents, yet ignored by most evaluated tools.

[8] The comparative evaluation of tools based on result list ranking is especially relevant in our case, given marked differences between the tools' result lists (less than 30% overlap in retrieved documents within the top-10 and top-100 of any tool pair).

[9] The standard included resources are intentionally kept scarce to keep the base selling price low.

On the interface side, support for reordering result lists based on any combination of metadata is considered useful, as well as dynamically generated summaries, taking into account the user's profile and information need. Short and high-quality summaries may save a user time in browsing and assessing the relevance of retrieved documents.

All evaluated tools implicitely assume that each document is independent from the others, where this is clearly not the case with BP's case reports. Support for logical document grouping (merging) would thus come in handy.

Lastly, all tools allow implementing user-based, document-level access policies, but none provides for fine-grained, function-driven rather than data-driven security mechanisms. The need for both kinds of security is governed by the 'need-to-know principle' as in force in the BP; officers may only access those parts of information they are entitled to for the sole purpose of carrying out their duties.[10]

3.2 Metadata Search

Purpose. The BP requires the tools to include a subset of their structured data (such as PV number, police zone, date range of fact, type of document, type of crime), and considers metadata search to be an essential functionality, leveraging free-text search.

Support. We found sufficient and similar support for the integration of metadata in all evaluated tools through the provision of connectors to the most common types of structured data sources (in particular ODBC gateways to RDBMS – relational databases).

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. These standard document attributes are e.g. used in search result lists, to give the user the gist of a document. Classification codes become available when indexing against one or more compiled taxonomies, and high-level document descriptors, such as entities or facts are included with the enabling of the associated tool services, if any (cf. infra).

Combinable search operators for the most common types of metadata (text, date, ordinal) are generally supported.

Quality. Due to the deterministic character of metadata search, qualitative evaluation is deemed irrelevant.

[10] Although not automatically enforceable, tools do provide for logging all transactions made, so that external (human) auditing remains possible.

Criteria. The automated detection of document encoding, type, and content language, and their inclusion as metadata, may prove an interesting feature. Not only can they be used to restrict the scope of search, they also allow for more flexibility in the organisation of documents on the file system. To explain this, we note that tools extensively make use of language-specific resources and natural language processing technologies, so that it is recommended to process documents by language separately. With no support for automated language detection, it is advisable to introduce high-level directories for each language, such that documents can be indexed (processed) separately by user-indicated language.

Other criteria are the ease of metadata integration and searchability, with sufficient support for metadata search operators.

Suggestions. We suggest yielding reliably generated metadata from text using information extraction technology to reduce the burdon of manual tagging and consistency problems. For the BP, given the description of case report outlines in Sect. 2.3, it should be possible to extract information fields (PV number, police zone,...) and identity information of victims, suspects, cars, objects, etc. in a reliable fashion.

Reordening the search result list based on any combination of metadata (e.g. police zone, date range, age group of a suspect) is a useful but not standard supported feature.

3.3 Classification

Purpose. Automated classification in the context of the BP can be used to organise the voluminous and heterogeneous document collection in manageable and content-specific units, reveal related and novel crimes based on their textual descriptions, or filter out relevant documents for divisions, investigators, or analysts specialised in different subject matters.

Support. Document retrieval tools typically offer classification as a standard extension to search and retrieval, employing manual rule-based taxonomy construction as the underlying technology. The expressivity of the rule language minimally entails classification based on literal occurences of keywords found in the text, and is in some tools maximally extended to the entire query language. This means one can create arbitrary search queries serving as definitions for content categories.

Data mining tools resort to the application of standard machine learning techniques on a chosen representation of the documents.[11] With supervised tech-

[11] Although an interesting application, the evaluated data mining tools do not offer document classification out-of-the-box; the available text mining and learning modules are merely used as essential components in a custom designed data flow model, leaving room for document representation and output dissemination.

niques, a classifier is trained on representative example documents for each category, and subsequently predicts the category of any newly presented document. Unsupervised techniques autonomously generate (a hierarchy of) clusters of related documents based on content similarity. This technique is also used in one document retrieval tool at search time (opposed to indexing time), dynamically clustering search results using salient noun and noun groups as representations of documents and displayed cluster labels.

Few tools allow for a hybrid approach, with the automatic generation of editable classification rules based on sets of representative example documents.

Quality. Implementing rule sets in practice turns out to be a painstaking activity that requires constant refinement and adaptation to cope with a constantly evolving stream of document content. Moreover, changes in the taxonomy's definition only become effective after reclassifying (often implying reindexing) the entire collection. This makes the rule-building approach practically hampered. Due to practical limitations, we were not able to draft a full rule set for each tool individually (using their distinct syntax) in order to seriously test their classification capabilities. Despite this fact, we can conclude - from studying the expressivity of the rule language provided - that there lies a significant gap between the lexical and semantic level of analysis, on the basis of which the tools, respectively humans are known to classify texts. This gap poses serious restrictions on what can be expected from automated classification by these tools. For example, when classifying by time of crime, a mention of '9:17am' in the text will not be recognised as belonging to the 'morning' category of the taxonomy unless explicitely coded for in the rule set, and even then it is not yet clear as to whether this mention pertains to the actual time of the crime, if any crime is described at all. Explicitely coding for all of this calls for a more advanced extension of the basic rule language provided by the tools, including world knowledge.

We tested supervised machine learning techniques (neural network and decision tree based) for a single-level taxonomy comprising three crime types: carjacking, pickpocketing, and money laundering. We used the *term vector* as traditional document representation ([8]). Even on this simple classification task, results never exceeded 82% precision and recall, which is clearly insufficient. Apparently, the traditional document representation as bag-of-words and adoption of default classifier settings (as preset by the tools) cannot adequately discriminate crime types in our document collection.

Criteria. Rule-based tools differ in the expressivity of the rule language, which may considerably facilitate the creation of the taxonomy's definition and contribute to its profoundness (e.g. use of fuzzy matching, lemmatisation, or linking with thesauri). As with free-text search, language support is also an important consideration.

Data mining tools often provide the same standard set of generic machine learning techniques (both symbolic and numeric, including decision trees, neural networks, and support vector machines). However, they differ to some extent in

their configurability, the provision of good default settings, as well as the speed of training, the support for different learning modes, such as batch or incremental.

Suggestions. For the automated classification of documents, we advocate a higher-level content analysis of texts that extends the traditional, lexical (word-based) content analysis as is merely performed by all evaluated tools. Like analysis might reveal that a hearing or case report 'fits' e.g. the scenario of a carjacking, a holdup, a declaration of stolen goods, etc. The analysis might be based on spatio-temporal relationships of facts described, with implicit or ambiguous scenario steps covered by context, common sense, and domain (world) knowledge. There are other cases in which documents can be more reliably classified using information units other than literal words. For example, within the BP, the main crime type can be easily derived from the PV number that is readily found in case reports using simple information extraction technology, such as regular expressions.

As with free-text search, crosslingual support is highly desired, as currently multiple rule sets have to be build and maintained in consistent manner for each language separately.

Lastly, classification is often considered a one-shot task in the evaluated tools, whereas a constantly changing content repository requires support for incremental learning, incremental taxonomy development, maintenance, and deployment, new topic signaling, and the like.

3.4 Named Entity Extraction

Purpose. Named entity recognition constitutes a basic operation in the structuring of texts. Its automation within the BP would e.g. tremendously aid operational analysts in the coding (schematisation) of criminal cases, sometimes covering hundreds of pages that otherwise would have to be skimmed manually for the discovery of entities of interest.

Support. Some core document retrieval tools venture in the domain of information extraction by allowing the automated discovery and classification of named entities in text, amongst which person names, organisations, locations, and time instances are most commonly supported. Text mining modules of data mining tools generally incorporate this feature too, whereas it forms the foundation of all tools truely profiled as information extraction tools.

All evaluated document retrieval tools largely rely on human editable and expandable dictionaries. Information extraction tools typically go one step further by providing human editable and expandable rule sets. A rule may be as simple as a regular expression, but may extend to incorporate lexical analysis of the source text. This way, the context of entity mentions may be used to trigger recognition or to determine their type (e.g. use of certain prepositions). We remain unsure as to which technology text mining modules resort to, but likely they use a combination of both.

Quality. We used standard measures of evaluation, namely precision, recall, and the combined F-measure ([8]) to assess the tools on a multilingual set of sample documents. We treated misalignment between extracted entity mentions and a golden standard of manually extracted mentions consistently in favour of the tools. This way, the extracted entity "South Africa" is equated with the full entity name "the Republic of South Africa", when the latter appears in the text.

Results show high precision on the most common entity types (persons, organisations, locations), up to 97%. Recall is very poor however, less than 50%. From these findings we assume the use of rather cautious dictionaries and/or rule sets, both limited in scope and their tolerance towards typographical, compositional, and other kinds of observed variations. This shortcoming is especially relevant given the noisiness of the texts in our collection. We also mention the problem of ambiguity, which gave rise to a number of errors, mostly when it comes to determining the entity type (e.g. locations or organisations named after persons).

Criteria. As techniques for entity extraction are type and language specific, the support and coverage of the base installation are important criteria. Other considerations are the provision of a clear, easily navigable user interface, sufficient import/export facilities with other tools (interoperability), and most importantly the quality of extraction (cf. supra).

Suggestions. None of the evaluated tools offers a learning approach to automated entity recognition, whereas the academic community has made much progress in this field ([9]). A line of research which is only tentatively pursued to date, is the extraction of entities within noisy texts ([10]).

3.5 Entity Linking

Purpose. Interrelating associated entities is performed within the BP to identify, visualise, and study criminal networks, as well as communication networks, transaction networks,... both in criminal investigations and for analysis purposes.

Support. Entity linking is the logical next step up to entity extraction, but has found no application in all evaluated document retrieval tools. The evaluated data mining tools provide some limited, generic visualisation modules that can e.g. be used to display networks of entities co-occuring in the same documents. One evaluated information extraction tool offers direct support for entity linking through the manual construction of extraction templates. Such a template denotes the presence of a fact of interest in the text, and can be arbitrarily composed out of keywords, lexical attributes (such as POS, identified entities) and structural markers (phrase, sentence, paragraph, text boundaries).

Quality. The quality of entity linking is influenced by the quality of entity extraction, on which it heavily depends. Further evaluation has not been pursued due to the rather limited capabilities offered by the tools in our selection.

Criteria. We refer to the same criteria as set out for named entity extraction (cf. supra).

Suggestions. Next to the manually constructed extraction templates, (weakly) supervised machine learning techniques could be implemented when having a representative (annotated) training corpus at one's disposal ([9]).

Being a desired and highly significant, yet advanced feature, coreference resolution has been scarcely addressed by any of the evaluated tools. In order to properly link entities within and across documents however (possibly of different languages), it is required that tools handle issues such as alias detection, disambiguation, anaphor resolution, temporal resolution, and the like. This becomes especially relevant with increasing joint efforts of security agencies sharing information, as mandated by the globalised nature of crime.

4 Related Work

In the past decade, many IT implementation projects have been conducted in collaboration with police forces throughout the world. Most noted are the COPLINK project of Chen et al. ([11]) in the state of Arizona, the CLEAR (Citizen Law Enforcement Analysis and Reporting) project in Chicaco, the FLINTS (Forensic Led Intelligence System) project developed since 1999 by West Midlands police under the auspices of R. M. Leary ([12]), and the OVER project of Oatley, Ewart en Zeleznikow ([13]), in association with West Midlands police since 2000. Most of these projects revolve around the centralization and consolidation of various digitized information sources. Applications range from information fusion, information sharing, improved availability (ubiquitousness) of information, to advanced exploitation for criminal analysis. In publications, only few attention is given to their (qualitative) assessment however. In the INFO-NS project ([2]), our primary purpose has been the evaluation of market-available commercial tools, for which only very limited studies exist.

Rijsbergen ([8]) has discussed evaluation techniques for measuring the performance of information retrieval tools. Related studies can be found from Lancaster ([14]), Cooper ([15]), and Ingwersen ([16]) on functional use assessment, relevance assessment, and quality evaluation, while the evaluation methodologies suggested by Elder and Abbot ([17]), Nakhaeizadeh, and Schnabl ([18]), Collier et al. ([19]) are notable. Cunningham et al. ([20]) has given a remarkable contribution to multilingual, multimedia information extraction ([21]).

5 Conclusions

In this paper we have conveyed our general findings with regard to the state-of-the-art of a selection of market-leading tools for the exploitation of unstructured information, as was the objective of the INFO-NS project carried out on behalf of the Belgian Police. In the areas of document retrieval, information extraction, and link analysis we have presented the support offered by these tools, reported crucial criteria and results on their qualitative assessment, and formulated recommendations on their possible improvement.

We believe a great potential for further research and tool development lies in the indexing and exploitation of multilingual, multimedia archives. This will undoubtedly push the envelope of current document repository and exploitation systems.

Acknowledgment

The authors would like to thank the Belgian Police for their interest and active collaboration, in particular Kris D'Hoore, Martine Pattyn and Paul Wouters. This work was supported by the Belgian Science Policy Office through their AGORA research programme ([22]). AG/01/101.

References

[1] Delphi Group, "Taxonomy and content classification - market milestone report," Online at http://www.delphigroup.com, Delphi Group, Ten Post Office Square, Boston, MA 02109-4603, Tech. Rep., 2002.

[2] N. Kumar, J. De Beer, J. Vanthienen, and M.-F. Moens, "Evaluation of intelligent exploitation tools for non-structured police information," in *Proceedings of the ICAIL Workshop on Data Mining, Information Extraction and Evidentiary Reasoning for Law Enforcement and Counter-terrorism*, 2005.

[3] J. De Beer and M.-F. Moens, "Rpref – a generalization of bpref towards graded relevance judgments," 2006.

[4] R. Russell and M. Odell, "Soundex," Patent 01 261 167, 1918.

[5] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, no. 4, pp. 845–848, 1965.

[6] 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.

[7] H. Blanken, T. Grabs, H.-G. Schek, R. Schenkel, and G. Weikum, *Intelligent Search on XML Data - Applications, Languages, Models, Implementations and Benchmarks*. Springer-Verlag, 2003.

[8] C. J. Van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworths London, 1979.

[9] M.-F. Moens, *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag, 2006.

[10] 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.

[11] R. V. Hauck, J. Schroeder, and H. Chen, "Coplink: Developing information sharing and criminal intelligence analysis technologies for law enforcement," in *Proceedings of the National Conference for Digital Government Research*, vol. 1, 2001, pp. 134–140.
[12] R. M. Leary, "The role of the national intelligence model and flints in improving police performance," Online at http://www.homeoffice.gov.uk/docs2/resconf2002daytwo.html.
[13] G. C. Oatley, B. W. Ewart, and J. Zeleznikow, "Decision support systems for police: Lessons from the application of data mining techniques to 'soft' forensic evidence," 2004.
[14] F. W. Lancaster, *Information Retrieval Systems: Characteristics, Testing and Evaluation*. Wiley, New York, 1968.
[15] W. S. Cooper, "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, no. 2, pp. 87–100, 1973.
[16] P. Ingwersen, *Information Retrieval Interaction*. London: Taylor Graham, 1992.
[17] J. F. Elder and D. W. Abbott, "A comparison of leading data mining tools," Tech. Rep., August 1998.
[18] G. Nakhaeizadeh and A. Schnabl, "Development of multi-criteria metrics for evaluation of data mining algorithms," in *Proceedings KDD-97*. AAAI Press, 1997.
[19] K. Collier, B. Carey, D. Sautter, and C. Marjaniemi, "A methodology for evaluating and selecting data mining software," in *Proceedings of the International Conference on System Sciences*, 1999.
[20] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks, "Multimedia indexing through multisource and multilingual information extraction: the MUMIS project," *Data & Knowledge Engineering*, vol. 48, no. 2, pp. 247–264, 2004.
[21] D. Maynard, H. Cunningham, and K. Bontcheva, "Multilingual adaptations of a reusable information extraction tool," in *Proceedings of the Demo Sessions of EACL*, 2003, pp. 219–222.
[22] Online at http://www.belspo.be/agora.