

Exhaustive Mining of Information from Unstructured Documents

Martin Soubotin, Sergei Soubotin
FreeText Software Technologies, Inc.
10065 Mallet Dr., Dayton
OH 45458, USA
<http://www.freetextsoftware.com/>
martin_soubotin@freetextsoftware.com

ABSTRACT

The paper describes some basic features of the project for developing a continuously running automatic procedure ensuring the most comprehensive, up to exhaustive, mining of information that is “hidden” in the unstructured text documents. By continuity of the procedure, we mean that information elements obtained at the earlier stages are used for subsequent processing. While the source documents are processed, a series of functionally distinct outputs is created, starting with extraction of metadata for processed documents, populating databases with extracted entities, identifying new essential facts, revealing hidden connections between people, organizations, etc., and finishing with building complex structures of extracted information elements. Examples of such structures are coherent free-text compilation report or a body of evidence supporting/invalidating assumptions.

Keywords: Unstructured Text, Information Mining, Named Entities, Information Synthesis, Implicit Information.

1. INTRODUCTION: EXHAUSTIVE INFORMATION MINING AS CONCEPT AND OBJECTIVE

For a long period, information retrieval has been largely identified with search for relevant documents. In the course of time, retrieval of textual information was extended beyond the search of whole documents, aiming at identifying relevant pieces of interest within them (passages, sentences, smaller text items). Pieces of interest can contain answers to questions posed in the query or provide an idea of a whole document content.

There is a clear trend towards more complete mining of information from unstructured documents.

Topic-relevant information items can be extracted applying the same query terms as in the documents

search. Also, same means of query expansion are involved, such as use of synonyms, derived words, frequently co-occurring terms.

Noteworthy however, that key words search limits the scope of output data to only information items containing query terms.

Such limits are not characteristic for ways of information retrieval that base on patterns approach. Patterns approach reveals and exploits specific features of expressions by which the search items are presented in written texts.

The respective technology - Information Extraction (IE) - became widely known since it has been tested at Message Understanding Conferences (MUC) organized by NIST (conducted from 1988 to 1998) [1, 2, 3].

Potentially, IE techniques are capable of providing a high degree of completeness and comprehensiveness of extracted subject-related information - if this information is contained in the processed documents. This can be ensured by using a predefined set of patterns covering all essential facets of the requested subject.

In this paper, we approach the task of achieving information comprehensiveness in the scope of a special focus: researching particular areas of activity.

One example area of activity is a business market segment viewed as a scene of events and relationships in which the players – market participants – are involved. Comprehensive information on this area should include in particular names of companies, their characteristic features, top executives, locations, important events, performed and planned actions, etc.

There exists a basic set of relationships – characteristics of the participants in any area of activity. Indeed, names of persons and organizations/groups, membership in an organization/group/community, their location, the leading/heading position in such, friendly/rival

relationships between actors are essential for researching a situation in any area of activity – professional, political, sportive, criminal etc.

The whole set of basic relationships can be recognized with a universal patterns set. In its turn, relationships that are specific for particular areas are handled with patterns, that use special vocabularies. In our experience, a special vocabulary for a given area of activity can be created in a relatively short period of time.

Thus, from this point of a focus on researching areas of activity, the technology exploiting patterns for comprehensive information mining can be regarded as largely domain-independent.

There is another, not least important, factor ensuring the most complete information mining from documents - deriving new information units from the already obtained.

The respective means are primarily represented by various procedures of transforming, combination and organizing extracted items. Information synthesis - a direction of text processing that currently attracts increasing attention - exploits a variety of such procedures [4].

Basing on the above-stated methodological prerequisites, we now can pose the objective of developing technology that would ensure comprehensiveness of the extracted information.

We suggest a term Exhaustive Mining of Information (EMI) as a name for such technology. The way of achieving comprehensiveness by EMI technology is based on:

- a) complete extraction of information items from unstructured documents;
- b) subsequent processing of the extracted items intended to transform and organize them in various ways, and infer new items from them.

The word “exhaustive” is used here to stress the aim at mining as much adequate information as possible, including latent information contained in the documents implicitly.

The background for the EMI technology project was provided by our longtime experiments with methods of search, extraction, and organization of textual information [5, 6, 7].

Within the framework of this project, we are developing technology for identification of basic relationships common to various areas of activity and demonstrate capabilities of this approach for competitive intelligence tasks.

2. EMI IN THE LIGHT OF CURRENT TRENDS IN INFORMATION RETRIEVAL

Satisfying Information Needs while Preventing Information Overload

The need for tools presenting a user with information contained in documents collections, was lately realized as an important objective by developers of the text processing software.

On the Text REtrieval Conference TREC-8, conducted by NIST in 1999, the Question Answering (QA) track was introduced for the first time, guided by the assumption “that users would usually prefer to be given the answer rather than find the answer themselves in a document” [8]. The goal of QA tracks is to “retrieve answers rather than documents in response to a question” [9].

Information Extraction techniques have demonstrated their practical usability for automatic filling the databases [10, 11].

Commercial solutions for named entities extraction are offered by a number of companies. Some of them also claim the ability to extract facts from the processed text, some other – to detect connections between entities [12, 13, 14].

Currently, there is growing evidence that the value of information derived from the documents may increase significantly due to procedures of categorization, combination and other forms of organizing and presenting the extracted items.

Ordering of extracted sentences/passages was recognized as a task originally in the field of multi-document summarization.

Barzilay et al. stated: “The problem of organizing information for multi-document summarization so that the generated summary is coherent has received relatively little attention. While sentence ordering for single document summarization can be determined from the ordering of sentences in the input article, this is not the case for multi-document summarization where summary sentences may be drawn from different input articles” [15].

Currently, issues of multi-document summarization are explored in the framework of Information Synthesis field.

As Amigo et al. put it, “From a Computational Linguistics point of view, Information Synthesis can be seen as a kind of topic-oriented, informative multi-document summarization, where the goal is to produce a single text as a compressed version of a set of documents with a minimum loss of relevant information. Unlike indicative summaries (which help to determine whether a document is relevant to a particular topic), informative summaries must be helpful to answer, for instance, factual questions about the topic.” [4]. The authors use the term “reports” to refer to the summaries produced in an Information Synthesis task, in order to distinguish them from other kinds of summaries.

By and large, the observed tendency for in-depth and more comprehensive mining of unstructured texts results in “more intensive exploitation of information sources already on hand” [16]. Thus, the more complete satisfying of users’ information needs can be achieved, while their information overload decreases.

Need for Providing Information at a Higher Level

There is a growing request for software products that make it possible to deal with complex problems.

Especial interest in such systems is expressed by the intelligence agencies. Their well-articulated request is for technology that “fosters and supports high-level exploration and interaction with information”, “provides information at a higher level, and frees analysts to focus on issues that matter” [16].

Not long ago, it seemed that results of such kind can be obtained only by humans. E.g., analysis of competing assumptions by means of testing their compatibility with the available information is considered by government analysts as one of the most important tasks, performed until now manually [17].

Our project is aimed in particular at computer-aided support for such complex information analysis tasks, that were up to now performed manually (see section 5).

Meanwhile, obtaining these kinds of results is possible only on the basis of reliable, accurate and complete extraction of entities connected by various types of relationships. We consider our patterns method as creating prerequisites for such extraction and as a basis for creating of more complex types of results.

We believe, that the trend towards deriving and delivering higher-level, more complex and concentrated information can be extended further. We are currently conducting more studies in this direction.

3. THE EMI PROJECT

EMI project is aimed at creation of continuous unified automatic procedure ensuring the most comprehensive and exhaustive mining of information that is “hidden” in unstructured text documents.

Continuity of such procedure presupposes the use of information elements obtained at the earlier stages for subsequent processing. Hence, in course of the successive automatic operations, a series of functionally distinct outputs is created, starting with metadata for processed documents, populating databases with extracted entities, finding new essential facts, revealing hidden connections between people, organizations, etc., and finishing with building complex structures of information elements (such as a body of evidence supporting/invalidating assumptions or a coherent compilation of all processed input information).

Under the EMI project, we have developed the system, that lets to obtain these kinds of results.

Though a user must be provided with results obtained at all levels, the high-level processing tasks are the subject of our especial attention.

Depending on the user’s purpose and the input information, the main application area of this technology can be defined as analytics (governmental and business-oriented) and research (for scientific, as well as educational purposes).

A request for exhaustive information mining can reflect the needs of a user interested in research/analysis of a problem or situation. For example, a user request can focus on a situation in some business activity area (business participants, their connections, their planned actions, essential events, etc.).

A user can query the Web via search engines or large off-line documents repositories. The query terms primarily determine the scope of documents that must be processed, while essential elements of information derived from the documents are mostly captured via patterns, not query terms.

4. THE USE OF SPECIFIC PATTERNS

Initial point in the sequence of technological procedures that largely determines the accuracy, recall and feasibility of information mining at all subsequent stages is identification of strings containing certain lexical-grammatical constructions. Essentially, such constructions include named entities and lexical expressions of relationships between them (often together with syntactic expressions, and some other text inclusions). Patterns, used by our system, capture the corresponding text strings and identify the informative components (entities and their relationships) inside these strings.

Identification of named entities (primarily, names of persons and companies, but also dates, places and so on) as parts of larger constructions ensures the reliability of their recognition. The EMI system reveals such types of relationships as “person - person”, “company - company”, “person - company”, “company - market area”, “company - planned/performed actions”, “company - event”, and some other. Mergers, takeovers, lawsuits, etc are recognized as essential events.

We first applied such type of patterns at the Question Answering track of the TREC conducted in 2001 and 2002 [6, 7]. Our conference reports (submitted at that time on behalf of the InsightSoft-M company) have initiated a useful discussion and gave rise to certain new ideas that extend this approach [18, 19, 20, 21]. Since that time, we have improved the original approach substantially, especially regarding the structure of

patterns, and use of rules for the analysis of pattern-matching strings.

5. BEYOND THE EXTRACTED ITEMS: OBTAINING AND ARRANGING FACTUAL STATEMENTS

Obtaining Statements

The first processing step of the EMI technology is extraction of entities and relationships. The output of the extraction step includes semantic metadata, data for populating databases, and lists of entities, such as specialists in a certain field. These data are used in further procedures resulting in concisely formulated factual statements, i.e. grammatically correct sentences expressing facts or assumptions, e.g., about involvement of identified people and companies in important events.

For example:

“Cisco Systems Inc. will soon announce its acquisition of Procket Networks.”

“PeopleSoft fired Conway as CEO, replacing him with Dave Duffield.”

“Lucent will collaborate with Juniper Networks Inc.”

Statements are usually obtained by grammatical transforming of strings containing the corresponding entities and relationships.

Verifying Assumptions

The factual statements obtained at the previous step can now be clustered and arranged in various ways for producing additional useful results.

As especially valuable, we consider such arrangement that presents statements in a sequential order, so that each of them semantically corresponds to the preceding one and finally to the initial statement. The initial statement is considered in such case as a verified hypothesis. Automatic arrangement of statements supporting a hypothesis is based on heuristic rules used to evaluate the degree of semantic similarity of sentences.

This procedure can be performed in relation to every possible future event or a planned action.

Example: among facts related to Oracle’s attempt to acquire PeopleSoft there are supporting the assumption *PeopleSoft will accept Oracle’s bid*:

“PeopleSoft shareholders accept Oracle’s bid conditions.”

“PeopleSoft shareholders tender their shares at Oracle’s bid conditions.”

“A majority of PeopleSoft stockholders tendered their shares.”

“Calpers, the biggest pension fund in the country, tendered its shares of PeopleSoft to Oracle.”

Items of evidence, refuting the above statement while

supporting an opposite one *PeopleSoft will reject Oracle’s bid*, can also be presented:

“PeopleSoft board unanimously agreed to oppose the Oracle’s bid”.

Normalized Statements

The factual statements and assumptions can then be transformed into a “normalized” form. In this procedure, their original words and phrases expressing relationships between entities are replaced with standard ones, denoting the corresponding category (e.g., “is allied with”, “has a partnership with”, “collaborates with” can be identified as belonging to the category of “partnership” and expressed by phrase: “is a partner of”. In this form, the extracted facts can be submitted to operations of logical arrangement.

Constructing Chains and Networks of Interrelated Entities

Such chains and networks can include companies connected by various types of connections, e.g., ownership, subsidiary, contractual relationships, etc., as well as by people. For example, a person may be regarded as a connecting link between companies as a member of the directors board of one company, a main shareholder of another and an owner of a third company.

The construction of networks makes it possible to uncover hidden (indirect) connections, e.g.:

“Ericsson is a partner of Cisco” → “Juniper is a partner of Ericsson” → “Lucent is a partner of Juniper”...

Here, each node is directly linked to the immediately preceding and indirectly – to the previous ones, allowing a user to see all companies indirectly connected to Cisco.

Reasoning

Statements can be used for deductive and inductive reasoning according to certain heuristic rules. The purpose is to obtain new statements the direct evidence for which is not present in the source texts. We can illustrate this by the following simple example:

Statement 1: *Oracle is a provider of software for payroll, accounting and human resources*

(obtained by transforming the extracted item “Oracle makes software used for payroll, accounting and human resources”).

Statement 2: *PeopleSoft is a provider of software for payroll, inventory and human resources*

(obtained from “PeopleSoft is the second-largest provider of software used to manage payroll, inventory and human resources”).

Conclusion: *Oracle and PeopleSoft compete as providers of software for payroll and human resources*

(derived statement obtained in the absence of formal evidence in the source texts).

6. EXTRACTING AND ORGANIZING TEXT PASSAGES

Passages as Information Units

The factual statements discussed above are useful, as far as the logical operations can be applied to them. However, information they contain, is present in corresponding passages of source documents in a more complete form, with necessary nuances and in a proper context. Such passages can be presented as information units per se.

Though text passages are not appropriate for logical procedures, they can be arranged in a discursive form, as a sequence where each succeeding passage complements and somehow specifies the preceding ones.

In order to be used by the ordering procedures, extracted passages must satisfy certain basic requirements. They are intended for being read outside of the full texts from which they have been extracted. Therefore, each extracted passage must be self-sufficient, i.e. understandable as it is, and express some "complete idea".

This way, the linguists usually define a "sentence"; but an idea cannot be completely expressed and understood if a sentence contains anaphora referring to preceding sentences. So, the latter should be considered as part of a self-sufficient passage. Besides, the meaning of a given sentence can be viewed as doubtful, or even negated from the point of the succeeding sentences. Thus, such succeeding sentences (that in original text are connected by means of certain syntactic elements) should also be included into the passage in order to express the complete idea.

Discursive Ordering of Passages

Our approach to discursive ordering of passages is implemented in our systems InformationCompiler and ResearchMagic [5]. The ordering is performed at two levels. At the first level, passages are grouped into sections. Each succeeding section covers a more specific, less general topic than the preceding ones. At the second level, passages within the sections are ordered. The principles of ordering rely on theories of coherence (specifically, on theories of thematic progression and "given-new" (theme-rheme) relationships of text units) [22, 23, 24].

The so ordered passages constitute the free-style report. See example of the free-style output (Fig. 1). Such compilations of information from the source documents can also be created as structured reports containing predefined sections. Both kinds of reports (free-style and structured) can be produced by InformationCompiler, so as by the currently developed EMI system.

Fig.1. Beginning of the Free-Style Report (Related to Oracle's Takeover of PeopleSoft)

LIST OF CONCEPTS

ORACLE
PEOPLESOFT
TAKEOVER
BID

REFERENCES

ORACLE

It will be up to the DOJ to show that that market segment would comprise only Oracle and SAP AG should the deal succeed.^[1]

The antitrust case hinged on a small slice of the overall market - a \$500 million niche focused on complex software applications that help manage financial and personnel departments for large companies, government agencies and schools. Because Oracle, PeopleSoft and Germany-based SAP dominate this market segment, antitrust regulators argued prices will rise and customer support will dwindle if one of the three rivals is eliminated.^[2]

The DOJ has argued that a merger between Oracle and PeopleSoft would drive up prices for HRM and FMS software because Oracle would have little incentive to give large customers discounts.^[3]

The DOJ's view that Oracle, PeopleSoft and SAP are the only suppliers which meet the needs of the large enterprise market for HRM and FMS software is "illogical and wrong". Other software suppliers produce HRM and FMS packages used by companies of varying sizes, including enterprises, Oracle argued.^[3]

7. OVERVIEW OF RESULTS OBTAINED BY EMI SYSTEM IN A CONTINUAL RUN INITIATED BY A USER'S QUERY

While a user is provided with results obtained at an initial level of processing, the system uses these results as input for second-level results, and so on. Altogether, six such levels can be distinguished. Below we describe the types of information available to a user at each level.

At level 1, entities and relationships are extracted and presented in three various forms:
as metadata for processed documents (names of people and companies, places, events, etc.);
as interrelated entities in structured form (tables and database records);

as lists of categorized entities (e.g., companies, executives, specialists, products, etc.).
 At level 2, factual statements are presented (e.g., “company A announced the intention to acquire company B”).
 At level 3, factual statements are arranged (e.g., in the form supporting an assumption about a planned action).
 At level 4, two types of results are created:
 networks of entities interrelated by categorized relationships (e.g., networks of directly and indirectly connected companies)
 deduced and induced statements (e.g., statements on rivalry of companies providing the same type of products/services).
 At level 5, a user gets documents passages containing information that corresponds to previously presented statements.
 At level 6, a user gets compilation reports built of text passages:
 structured reports (with predefined sections) on previously mentioned people and organizations
 coherent free-style report containing the compilation of all processed documents.

8. CONCLUSION

The tendency towards comprehensive mining of text documents as well as increasing demand for providing information at a higher level face the developers in front of a new task: creating technologies aimed at a most comprehensive mining of information hidden in documents and presenting it to the user in the most organized form. If documents covering an area of user’s interest are addressed, not only all the pertinent data must be obtained, but also - information derived from combining and judicious arranging of these data elements. The unstructured texts are rich in implicit information all of that should be made explicit. The EMI technology described here represents, we believe, a step in this direction.

9. REFERENCES

- [1] Wendy Lehnert, Beth Sundheim, An evaluation of text analysis technologies, **AI Magazine**, v.12, n.3, pp. 81-94, 1991.
- [2] S. B. Huffman, Learning information extraction patterns from examples, **Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing**, pp.246-260, 1996.
- [3] J. Cowie, W. Lehnert, Information extraction, **Communications of the ACM**, v.39, n.1, 1996, pp. 80-91.
- [4] E. Amigo, J. Gonzalo, V. Peinado, A. Penas, F. Verdejo. An Empirical Study of Information Synthesis Tasks, nlp.uned.es/pergamus/pubs/articuloACL2004.pdf
- [5] <http://www.freetextsoftware.com>
- [6] M.M. Soubbotin and S.M. Soubbotin. 2002. Use of patterns for detection of likely answer strings: A systematic approach. In: **Proceedings of the TREC-2002 Conference**.
- [7] M.M Soubbotin, S.M. Soubbotin: Patterns of Potential Answer Expressions as Clues to the Right Answer, **Proceedings of TREC-10 Conference** . Gaithersburg, MD, 2001.
- [8] E. M. Voorhees. The TREC-8 Question Answering Track Report. In: E. M. Voorhees and D. K. Harman, editors, **Proceedings of the Eighth Text Retrieval Conference**, pp. 77–82, 1999.
- [9] E. M. Voorhees. Overview of the TREC-9 Question Answering Track. In: E. M. Voorhees and D. K. Harman, editors, **Proceedings of the Ninth Text Retrieval Conference**, pp. 71–79, 2000.
- [10] R. Grishman. Information extraction: Techniques and challenges. **Lecture Notes in Computer Science**, 1299:10-27, 1997.
- [11] A. Bia, R. Munoz, Information Extraction to feed Digital Library Databases, <http://www.cervantesvirtual.com/research/articles/sepln00.pdf>
- [12] Welcome to Inxight Software, Inc.. <http://www.inxight.com>
- [13] Stratify - Unstructured Data Management Solutions <http://www.stratify.com/>
- [14] ClearForest: Text-Driven Business Intelligence, <http://www.clearforest.com/>
- [15] R. Barzilay, N. Elhadad, K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. **Journal of Artificial Intelligence Research**, v. 17, 2002, pp. 35-55
- [16] ARDA Home Page, <http://www.ic-arda.org/>
- [17] R. J. Heuer. **Psychology of Intelligence Analysis**. Center for the Study of Intelligence, Central Intelligence Agency, 1999.
- [18] E. M. Voorhees. Overview of the TREC 2002 Question Answering Track. In E. M. Voorhees and L. P. Buckland, editors, **Proceedings of the Eleventh Text Retrieval Conference**, 2002.
- [19] Deepak Ravichandran and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**, Philadelphia, 2002, pp. 41-47.
- [20] G. B. Anaya and L Kosseim. Generation of natural responses through syntactic patterns. **TALN**, 2003.
- [21] Valentin Jijkoun and Maarten de Rijke Jori Mur. Information Extraction for Question Answering: Improving Recall Through Syntactic Patterns, <http://www.text-mining.org/>
- [22] Danes, F. (ed.) 1974. **Papers on Functional Sentence Perspective**. The Hague: Mouton.
- [23] Fries, Peter H. 1983. On the Status of Theme in English: Arguments from Discourse. In: Petöfi, J. S. and E. Sozer. (eds.). **Micro and Macro Connexity of Texts**. Hamburg: Buske: pp. 116-152.
- [24] Petöfi, Janos (ed.). 1988. **Text and Discourse Constitution**. Berlin: Gruyter.