USING TEXT CLUSTERING FOR INTELLIGENCE CLASSIFICATION

Tomas Berg, Christian Mårtenson, Pontus Svenson[1]

Swedish Defence Reseach Agency
Department of Decision Support Systems
SE 164 90 Stockholm, Sweden
ponsve@foi.se

Abstract
In this paper, we discuss how text mining methods could be used in a mixed-initiative interaction approach to intelligence analysis. We describe how simple methods from text mining can be used to help intelligence analysts determine where a specific report or analysis fits into the knowledge base (KB), i.e., how it should be classified and which, if any, other documents in the KB it should be linked to. The method works by comparing the vector space model representation of the new information document with those of all documents previously stored in the knowledge base. Those documents that are sufficiently similar to the new piece of information are displayed to the user, who can then choose to place links between them. Using a computer tool such as the one suggested here potentially allows the analyst to spend more time analyzing intelligence reports rather than searching for and classifying them. In previous work, we have discussed how the MilWiki, an improved implementation of the open-source MediaWiki system, could be used as a knowledge base for military purposes. To illustrate the text classification method described in this paper, it has been implemented for MilWiki. To simulate new pieces of information, the prototype allows the user to download articles from the Wikipedia. The method, as well as the collaborative work process used in a wiki, could be implemented in any content management systems. In addition to describing the text classification method, we also give a brief introduction to text mining and the vector space model of documents.

Introduction

Military commanders and analysts need decision support to help them attain situational awareness and plan their future actions. The decision support can consist of several things: human support in a reachback function, a validated methodology for how to process data and information, and computer tools that aid the process. The computer tools must be powerful enough to add value to the intelligence analysis process, yet simple enough to use so that they do not add unnecessarily to the cognitive burden of the users. Put shortly, the tools must be useworthy, not only useable or powerful.

[1] Corresponding author

Historically, most technical decision support research has focused on quickly processing information from sensors to construct situation pictures and impact assessments. However, the increasing use of computers both by blue force soldiers and by the people in the areas we are operating in means that more and more text documents will be produced that contain information that is useful for achieving situational awareness. There is thus a need for building decision support systems that integrate handling of textual reports with the handling of structured reports from sensors.

In the areas where the future Nordic Battle Group and other Swedish forces will be operating, there will be many different actors present. Only some of these will be openly hostile towards us and the peace-keeping or peace-enforcing mission that we have. The attitude of the different actors towards us will also change rapidly, perhaps from day to day or week to week. To keep track of the attitude that one group has towards us, it is useful to look at the written material that the group produces. This material could be in traditional newspapers, on leaflets or posters distributed in the city, or on web sites. To be able to analyze these reports quickly enough to be useful, we need text mining tools.

The example application presented here, WikiImporter, was motivated by a question we received when presenting our earlier work on the MilWiki KB concept to a Swedish intelligence officer. The question was how to easily transfer information that is contained in one KB into another. How can we check to make sure that the information is not already present in the KB? Is it possible to automatically see what documents in the KB that the new document should refer to? The example that the officer had in mind was to transfer information from a KB for the civilian-military cooperation function to the intelligence function of the battle group headquarters, but the same issue arises also in many other cases, for instance when information should be transferred from an unclassified KB into a classified, or when debriefing information from a mission should be transferred into a lessons learned system. The tool described in this paper describes how this problem could be partially solved.

We start by giving an overview of text mining, introducing some of the most important concepts from the field of research. Next, the WikiImporter is presented, a simple application that helps users to import document from one knowledge base into another. This is followed by a brief list of other useful application of text mining techniques. We conclude with some conclusions.

Text mining

In the intelligence process in Operations Other Than War (OOTW), it is necessary to analyze vast amounts of data. In addition to data that comes from the large numbers of sensors envisioned in the Network-based Defense concept, for OOTW it is also vitally important to consider text-based sources of intelligence, such as media reports and material published by NGOs. If we look a few years into the future, it is probable that people living in countries where the Swedish force is operating will use blogs and web sites to describe their lives[2]. Information that can be gathered from

[2] This idea is not as far-fetched as it might seem at first sight. Today in Iraq, there are a considerable number of blogs that describe how ordinary people live their lives.

these sources can be useful but of course also biased, and must hence be thoroughly analyzed and compared to other sources before decisions can be based upon them.
In OOTW, it is also necessary to handle all text reports made by soldiers who are out patrolling or on surveillance missions. These reports, while probably still conforming to the standard 7S-format, will also include free-text. Intelligence analysis in OOTW situations is considerably more difficult than in traditional warfare. Often, we will not even know what specific questions to ask the soldiers who are going on patrol or reconnaissance missions. Instead, they should be encouraged to make note of "everything" that they see, and provide as detailed as possible accounts of their observations while on patrol [1].
In order to handle these large amounts of data, computer tools that allow the analysts to focus their attention on the actual thinking required to produce situational assessments are needed. The analysts should not need to first sort through a large number of reports before they can begin their analysis. This boring and unproductive work should instead, as much as possible, be performed by computers. By letting the computers do the boring, non-creative work, the analysts are given more time spend on creative thinking and analysis.
Other organizations and enterprises that produce large amounts of data have already drawn the same conclusion. There exist many business intelligence applications that exploit the large data repositories to deliver actionable knowledge to decision makers. Traditionally, these tools have focused on the data structured in the enterprise databases, trying to discover information that explains a situation or can predict the next. This procedure of extracting valuable patterns from structured data is called *data mining*.
However, the majority of enterprise data is *not* structured in the sense that makes it readily available to data mining tools. At least 80 % of the information is kept as text documents with little or no structured metadata at all [4, 7]. Nevertheless, useful information can be extracted from these documents. The technology that is used to do this is called text mining [6], and is a hot research topic today.
Techniques for text mining are closely related to those of data mining. A large part of text mining consists of transforming the unstructured documents into structured tables (spreadsheets) so that data mining techniques can be applied.

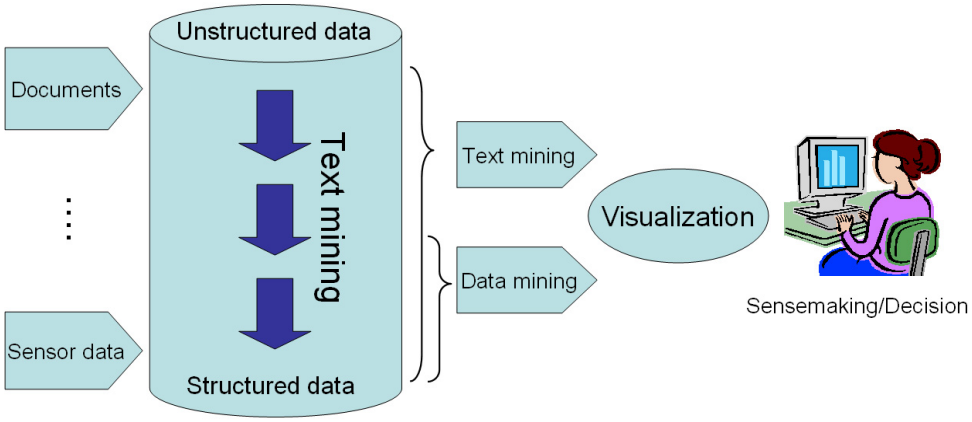
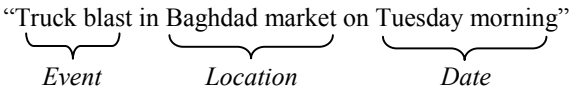


Figure 1. Simplified view of the relations between data and text mining. See text.

In figure 1, we attempt to show the relations and differences between data and text mining. The database (large cylinder) contains information with varying degrees of structure, ranging from completely unstructured documents (e.g., text) at the top to structured content at the bottom. Data is fed into the database by a wide variety of data sources. Within the database, text mining techniques are used to add structure to the unstructured content. The structure added can be of many different kinds: the least structure that needs to be added in order for the information to be useful is to index the document and update the document vector model used for searching. More sophisticated structuring methods include entity extraction, see below. The goal of information analysis processes is to present results to a human user who uses them to make sense of a situation and make a decision. Both data mining and text mining can be used to extract the information that is to be visualized to the user. The border between the two techniques is fuzzy. The more structure one has in the information (i.e. the lower one gets in the picture), the more relevant it is to talk about data mining instead of text mining. The picture is not meant to be seen as a process description, but rather provides a way categorising the different functions that are needed in the decision support system. Note in particular that we do not explicitly show the feedback loops that are a necessary part of the intelligence analysis process. Upon receiving the results from the data and text mining procedures, the user will not (except in the most trivial cases) immediately attain situational awareness. Instead, it will be necessary to iteratively refine the queries posed to the data base, and perhaps also to gather more information from sensors and other sources. The intelligence analysis process is a loop, not a line.
As noted above, there are a number of different methods for performing information structuring using text mining. They can roughly be divided into two categories, one drawing on linguistics and the other on statistics.

The linguistic approach utilizes methods from Natural Language Processing to extract entities and relations from documents and give them well-defined meaning by mapping them on a semantic structure, such as an ontology. The extraction is

performed by parsing the document while trying to recognize patterns that can reveal if a word or phrase can be classified as a known entity or relational class. The pattern recognition can be based on syntactical or grammatical rules as well as on lexicons and statistically derived patterns. The extracted entities can be for instance persons, organizations, dates, money, locations or events.



Extracting entities means that parts of the information contained in a document turns into structured information, possibly stored as metadata tags. Although the extraction process will not be one hundred percent accurate unless continuously supervised by a human, the automatically generated metadata will immediately be available for machine interpretation. The metadata can also be used to summarize the document for a human operator, or can be used to guide the analyst in determining what parts of the document that it is most relevant to look at. This will allow large amounts of information to be processed and interpreted in fractions of the time needed to perform a completely manual analysis.

Even more information can be extracted by analyzing the relationships between the extracted entities. This will lead to entity-relationship diagrams such as the one in figure 2.

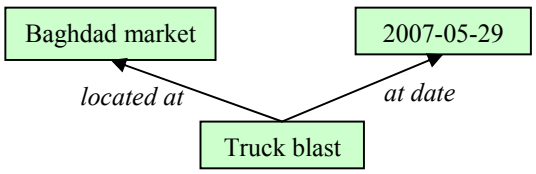


Figure 2. Entity-relationship diagram connecting extracted entities with semantic links.

Such diagrams could form the basis for *conceptual maps* that help the human in the sense making process. Entity diagrams that relate individuals and organizations to each other could also be input to social network analysis programs.

The statistical approach is based on calculating frequencies (the number of occurrences) of the different words in each document, and then trying to find patterns in their frequency profiles. The frequency profiles are usually represented by column vectors in a term-document matrix, A, where element a_{ij} is the frequency of word i in document j (figure 3). This so called *vector space model* is suitable for performing a number of fundamental mining tasks, such as information retrieval, clustering and classification.

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	…
abort	0	0	0	0	0	0	0	0	…
absolute	0	0	**1**	0	0	0	0	0	…
absorbed	0	0	0	0	0	0	0	0	…

acceptability	0	0	0	0	0	0	0	0	…
…	…	…	…	…	…	…	…	…	…
zone	0	0	0	0	0	0	**1**	0	…

Figure 3. The figure shows a term-document matrix representing term occurrences in documents. 'Absolute' occurs in document D3 and 'zone' in document D7. The size of a term-document matrix is typically very large, possibly containing hundreds of thousands of words and millions of documents. Luckily they are also very sparse (i.e., they contain many zeros) which makes computations and storage feasible.

Text Retrieval is the science of searching for information in document collections. It is a subfield of Information Retrieval which deals with search in all sorts of collections, ranging from databases with numeric data to different types of media repositories. Text retrieval applications mostly consist of web and enterprise search engines (e.g. Google, Yahoo!, Fast S&T, Autonomy). The key technology is text mining which is used to determine the proximity between a search query and the documents in a repository. The query is represented as a document vector and the distance to the other vectors in the term-document matrix can be determined by for instance calculating the cosine of the angle between the vectors. The most similar document vectors get the highest rank in the results list. The simple importer example application presented below makes use of such search queries to find relevant documents.

Clustering is a method for revealing hidden structures in large data sets by automatically grouping objects with similar features. In text mining, clustering gives rise to groups of documents with similar content. The similarity is based on the distances in the vector space model as described above. Clustering can be helpful in getting a quick overview of large repositories. In information retrieval it is used for grouping search results, allowing the user to narrow the search to the most relevant cluster.

Text categorization (or document classification) is a predictive mining technique that exploits machine learning to automatically put documents in different predetermined classes. The learner is presented with a number of correctly classified documents and is then with some accuracy able to predict the classes of new documents. The classification is based on how similar documents have been classified previously, again using proximity measures in the vector space model.

Many examples of the power of statistical methods for information retrieval and natural language processing can be found by studying the products of Google. Google search works by calculating a document vector model for the entire web. Ranking of the results presented for a user query is then performed by considering also the network structure of the web: a page that is linked to by many important pages is itself important, and is thus presented earlier in the search results page. Google also has a free translator service, which uses statistical methods on repositories that contain the same documents in many different languages to provide translations of user-input text.

It is possible to make many improvements on the rather simple method for determining document similarity presented above. One very useful technique, called latent semantic indexing, is to make a mathematical transformation of the matrix

used to store the vector space model. In this transformed matrix, a simple approximation which corresponds to doing a principal component analysis and only retaining the largest eigenvalues, is made. In the transformed version, the words that correspond to the y-axis of the matrix are transformed into linear combinations of words that are termed “concepts”. Each concept represents terms that are used in similar way in the document repository. The concepts can sometimes, but certainly not always, correspond to synonyms, and are thus found automatically. After transforming back into the original representations, the matrix will contain traces of these concepts, which can be used to decide that two documents are similar if they refer to the same concepts, even if they do not use the same explicit words for the concepts.

The Wiki-Importer application

To illustrate some of the benefits that text mining techniques could give to the military, we have implemented a simple application called WikiImporter. The idea behind this prototype is to demonstrate how text clustering could be used to help analysts transfer knowledge from one knowledge base system into another. The WikiImporter application allows a user to import a page from one Wiki installation into another. We chose to use wikis to demonstrate the concept because we have previously used it as an example KB system [2].

A wiki is a system that enables users to collaboratively create and edit web content directly, using a web browser. The content is usually stored in a relational database, which is run on one or more web servers. Wikis combine search mechanisms with a history mechanism that provides version control and allows authorised users to undo changes made by mistake. Most Wikis do not provide support for user authentication beyond a simple password system or integrate access rights handling into the software. There are however exceptions, and a wiki that is used for military purposes must of course have very strict handling of access rights. To make full use of a KB, it is important to have relevant *links* between the different documents that are stored in the KB.

WikiImporter works by using the “More Like This” function of the Lucene search toolkit. This function searches a document database and uses text clustering to find those documents that are similar to a provided target document. The documents returned by the “More Like This” function are the ones presented to the user as possible candidates for linking to the imported document. Figure 4 shows the user interface of the concept prototype. Here, we have chosen to import an article about the “Saab Lansen” aircraft from the WikiPedia into the MilWiki. As can be seen on the right, a number of similar documents have been found, among them the MilWiki page for the “Saab Viggen” aircraft. The user is told that there is no perfect match (as there could be if the information in the imported document was already present under a different name in the KB), and can select to create a page for “Saab Lansen” in the KB and to link that page to any of the returned list of similar documents.

Figure 4. This figure shows the user interface of the concept prototype that was developed.

Several things are worth pointing out after this example. What is presented to the user is not a finished result, but rather suggestions. It is not possible to construct computer tools that determine automatically whether a found page really should be linked to the newly imported one. This decision must be taken by the human operator. The application implemented is thus an example of a Mixed-Initiative Interaction (MII) [3] tool, where the best sides of both the human and the computer are combined. Another similarity to MII is that the list of similar pages could equally well have been supplied by a human.

Other applications of text mining

The application presented above is rather simple, but nevertheless useful. There are a number of other possible uses of text mining techniques that are useful for decision support systems. Here, we very briefly describe just a small sampling of possible applications.

In other work [5], we have presented a tool for impact assessment that makes use of so-called indicators to connect incoming observations to events that pose a threat to us. In the current implementation of this tool, indicators are set manually for the reports. By using techniques from text mining, it should be possible to suggest possible indicators to the human operator. In this application, it is likely that the linguistic approach will be most useful for 7S format reports, while the statistical

approach will prove fruitful for helping the user to analyse longer articles, from, e.g., newspapers.

An interesting line of research in text mining deals with fusing several different descriptions of a set of events into a single storyline. By structuring the accounts given by different people and time-ordering them, it is possible both to get a more complete account than if we just considered one of the witnesses, and to determine parts of the storyline where there are discrepancies between the different accounts. The resulting storyline could be used for building an organizational memory of what happened during different missions, which is useful for training for future missions. This technique could also potentially be useful for war-crime investigations, or, more generally, in other kinds of criminal investigations. Today's state-of-the-art speech recognition software is good enough that it is possible to obtain transcriptions of evidence given by different suspects and witnesses in crime investigations. Performing entity extraction on the resulting texts and summarising them would allow police officers to quickly see patterns and inconsistencies in the evidence.

Text summarization could be used to provide analysts with summaries of what is reported from the area of operations. It is of course important to not only provide the analysts with summaries that describe what the different sources agree on. The parts of the events where the sources do not agree must also be highlighted for the user. The summaries computed during a mission should also be influenced by the current situation: e.g., if a specific war lord is currently very active, the summaries should be more detailed about their activities. The summaries that are computed should not only be text-based, but also include concept maps generated by doing entity and reltion extraction on the documents. Text summarisation algorithms work in two different ways. One approach is to select the most relevant sentences from the text and display them to the user. For instance, a very simple algorithm of this type could be to always select the first sentence in each chapter of the document, and display these as a summary. A slightly more sophisticated algorithm would chose those sentences that contain the words that are most characteristic of the document. These words would be determined by representing the document using the vector space model described above, and comparing the vector space representation with those of a large ensemble of background documents, in much the same way as the Wiki Importer application does. The second approach to summarisation is to select a consecutive block of text from the document and let it represent the document. This approach is useful if the documents one wishes to summarise are written as newspaper articles often are, but can also lead to very misleading results.

The ideas presented here are of various degrees of sophistication. The difficulty with which they could be implemented also varies. In particular, much research is still needed in the area of text summarisation.

Conclusions

In conclusion, we presented a simple tool that can help an intelligence analyst to import reports from one intelligence KB to another. The tool compares the new intelligence report with the ones already present in the KB, and shows the "closest" reports to the user. The user can then choose to add links between the selected old reports and the new one. We emphasize that the tool is not meant to be fully

automatic: it does not have enough precision for that. Instead, it is meant to help the user by cutting down on the number of possible links that it is necessary to consider.

The tool frees the user/intelligence officer from some of the non-creative work that is involved in the intelligence analysis process and thus allows the analyst to spend more time actually analyzing. We also described some of the basics of text mining, and presented a list of other possible uses of text mining technology that could be used to achieve situational awareness. The importer tool was implemented for the MilWiki [2] prototype KB, but the concept is of course not tied to this specific implementation of a KB.

The presented tool does not aim to be very sophisticated, and does not require the users to change their process for analysing intelligence. It gives suggestions to the user in much the same way as a human does. Like with advice from other humans, the user should not always trust the results presented by the program, but rather use them as a help to cut down on the number of alternatives that they need to consider. The tool serves as a prototype application that implements mixed-initiative interaction using rather straight-forward techniques from text mining. There is room for considerably more research on how to construct more advanced intelligence analysis applications.

References

[1] Folke Andersson Elisabeth André Turlind, Eric Scöberg, Åke Wiss, Å., "Framtida konflikters karaktär: delrapport 3" [The character of future conflict: report 3], Technical report FOI-R—0684—SE, 2002

[2] Mikael Brännström, Christian Mårtenson, "Enhancing situational awareness by exploiting wiki technology", Proceedings of Conference on Civil-Military Cooperation (CIMI) 2006

[3] Katarina Johansson, Christian Mårtenson, "Mixed-initiative-interaction för militära beslutsstödssystem" [Mixed-Initiative Interaction for military decision support systems], Technical report FOI-MEMO—1864, 2006

[4] Cathleen Moore, "Diving into Data," *InfoWorld* (October 25, 2002), *http://www.infoworld.com/article/02/10/25/021028feundata_1.html*

[5] Pontus Svenson, Tomas Berg, Pontus Hörling, Michael Malm, Christian Mårtenson, "Using the impact matrix for predictive situational awareness", In Proceedings of the 10th International Conference on Information Fusion, 2007

[6] Sholom Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer, 1:st ed. 2004. (ISBN 0387954333).

[7] Colin White, "Consolidating, Accessing and Analyzing Unstructured Data", Business Intelligence Network (December 12, 2005), http://www.b-eye-network.com/view/2098