

VISUALIZATION OF NEWS ARTICLES

Marko Grobelnik, Dunja Mladenic
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: marko.grobelnik@ijs.si

ABSTRACT

This paper presents a system for visualization of large amounts of new stories. In the first phase, the new stories are preprocessed for the purpose of name-entity extraction. Next, a graph of relationships between the extracted name entities is created, where each name entity represents one vertex in the graph and two name entities are connected if they appear in the same document. The graph of entities is presented as a local neighborhood enriched with additional contextual information in the form of characteristic keywords and related name entities connected to the entity in the focus. Operations for browsing a graph are implemented to be efficient enabling quick capturing of large amounts of information present in the original text.

1 INTRODUCTION

Text visualization is an area having the main goal to present textual contents of one or many documents in a visual form. The intention of producing visualization of the textual contents is mainly to create graphical form of content summary on different levels of abstraction.

In general, we could say that ideas used in text visualization algorithms come primarily from data analysis research areas (such as statistics, machine learning, data mining) [1, 2, 3, 4] where data visualization play important role as a key technique for showing the data and results of analytic methods. Textual data is in this respect just another type of data with its specific properties which need to be taken into account when visualizing it. Main characteristics relevant for text visualization are [5]:

- High data dimensionality when using typical bag-of-words representation, where each word and each phrase represents one dimension in the data space.
- High redundancy, meaning that many dimensions can be easily merged into one dimension without losing much information. This is caused by the two properties of words, namely synonymy (different surface word form having the same meaning – e.g. singer, vocalist) and hyponymy (one word denotes a subclass of another – e.g. breakfast, is a subclass of a meal)
- Ambiguity between words in the cases where the same surface form of the word has different meanings (homonymy – e.g. the word ‘bank’ can mean ‘river bank’ or ‘financial institution’) or in the cases where the same form has related meaning (polysemy – e.g. ‘bank’ can mean ‘blood bank’ or ‘financial institution’)

- Frequency of words (and phrases) follows power distribution, meaning that we deal with small number of very frequent words and high number of infrequent words. Having this in mind, we need to use appropriate weighting schemas (e.g., most popular being TFIDF) to normalize importance of the words to be able to work with the standard data analytic techniques.

Furthermore, when talking about text visualization we also need to be aware of the type of text we are dealing with. Namely, different document types have different characteristics which need to be considered when designing an efficient text visualization mechanism. Some examples of such different types of textual data are: Web documents (being typically short, having linkage structure and additional formatting information), e-mails and news-group postings (short documents with specific internal structure, appearing in content threads and using specific language), customer reports, chat rooms discussions, literature, legal documents, technical text, news stories etc.

In this paper we are dealing with news stories. Specifically, we are presenting a system for preprocessing and visualizing large amounts of documents coming from a news wire. In general, news stories are special type of text having most often the following properties:

- shorter documents,
- written by professionals,
- low number of mistakes,
- having good rhetorical structure,
- rich information about people, companies, places, etc.,
- single documents being pieces of larger stories spanning over several documents.

Our approach takes into account the above properties giving a special emphasis on the last two items namely, named objects (such as people, companies, and places) and the context they are appearing in.

In the following sections we present related work, sample news articles corpus, design and architecture of the system, name entity extraction, keyword extraction, browsing and visualization user interface and discussion at the end.

2 RELATED WORK

Wider area of the work presented in this paper fits in the areas of data visualization [3] and in particular in the areas of text visualization [6] and more recently, developments of semantic web and visualization of ontologies and other knowledge structures [1].

In particular, in this paper we are dealing with visualization and browsing of news stories which require special treatment. In the literature there are not many published works on this specific subtopic. Most prominent is the overview publication from MITRE team [7] giving good overview over the approaches for visualization of different document types, including news stories. Their goals are similar to the work presented here, but the actual approach is quite different. Their publication appeared also at [8] together with some other interesting approaches for document visualization.

Another approach for visualizing trends in news documents is the system ThemeRiver [9] developed at Pacific Northwest National Laboratory together with many other interesting approaches for information text visualization [10]. ThemeRiver in particular is specialized for analyzing and visualizing trends in news stories over time, enabling efficient detection of trends in the vocabulary used in the texts. Among others, we would also like to mention work of the authors on visualization of large text corpora [11] presented at “TELRI - Information in Corpora” workshop which directly precedes this work.

3 SAMPLE NEWS CORPUS

The functionality of our approach is presented here on a corpus of news articles from “ACM Technology News” service at <http://www.acm.org/technews/archives.html>. The corpus includes general news from the most areas of Information Technology (December 1999 on). It includes over 11.000 article summaries of the length 200-400 words.

4 DESIGN AND ARCHITECTURE OF THE SYSTEM

The main goal, when designing the system called “Contexter”, was to help expert and semi-expert users (such as analysts, journalists, social scientists, experienced web surfers) to get an efficient and quick understanding of large corpus of general news stories at different levels of abstraction. This is to be achieved by several means:

- by showing relationships between entities appearing within documents,
- by calculating and showing contexts within which the entities appear either individually or in combination with other entities,
- by using several types of visualizations,
- by efficient and responsive graphical user interface with easy moving from abstract to detailed information.

One of the fundamental design assumptions is that most of the relevant information is centered around the entities mentioned within the documents. In our context, entities can be names of people, names of companies and other institutions, geographical names and places, product names, etc. An additional property of entities is that they serve as connectors between different documents forming longer threads of stories which are not explicitly noted with typical news corpora (usually such information is not present in

meta-data of news articles). Based on these observations, our basic representation of documents within the news corpus is three-fold: (1) plain text as originally provided, (2) bag-of-words representation of the text, (3) representation by a set of name -entities:

1. Plain news text as written by the authors. This representation is used exclusively for showing the document contents to the user in the situations when the user comes to the point where it requests full textual information. This representation offers lowest level of content abstraction.
2. Bag-of-words representation using some kind of weighting schema (in our case TFIDF). In this case we still include most of the words appearing in the original text – we just delete the stop-words (non-informative functional words), perform stemming (unifying different surface forms for the same words), pre-calculate phrases (frequent and significant consecutive sequences of several words), and the most important, ignore the order of the words (for the purpose of more efficient computation). The goal of this representation is to efficiently calculate contexts in the form of keyword lists to allow for a higher abstraction of the contents compared to the plain text.
3. Set of name -entities appearing within the article. In our case we use variant of relatively standard name-entity extraction algorithm based on word capitalization (primary candidates for the name -entities are the words starting with capital letter) with additional mechanism for name consolidation (detecting that e.g. ‘Bill Clinton’=‘President Clinton’=‘Clinton’). This representation in our case offers the highest abstraction level for an individual document. Because of its structured nature (e.g. names are consolidated on the level of the whole news corpus) it serves as a connecting level between different documents.

On the input to the system we get a set of documents representing news articles. We have no special assumptions on the form, structure and meta-data within the documents – main element is textual part of the documents which is further processed. Next, the documents are preprocessed in two different ways. First, the text is cleaned and bag-of-words representation is created, and next, the name -entities are extracted. All the documents are stored in three different representations (as already described: plain text, bag-of-words and name -entities) in the database which is then used by the client software using efficient graphical user interface described in the following sections.

5 NAMED ENTITY EXTRACTION

Information extraction and specifically name -entity-extraction [4] are one of the most popular areas of text mining. The main function is to convert parts of unstructured textual data into structured form which enables

to use standard data analytic methods available in data mining and statistical packages (e.g. SAS and SPSS use this kind of approach). There are three main approaches when extracting useful pieces of information from text: (1) manual extraction rules, (2) automatically generated rules with machine learning methods, and (3) hybrid methods combining the two approaches. In everyday practice the approach with manual rules seems to be most effective and most often used – machine learning methods give good results on datasets with lack of domain knowledge, but still automatically generated rules usually need human corrections and additions to be practically useful. In general, for controlled corpora, initial investment to get good results with manual rules seems most price-performance effective.

In our case, name entity extraction algorithm is based on one of the most typical heuristic approaches – on word capitalization. This approach usually gives good results on high quality texts and introduces low overhead in terms of computational efficiency and additional tuning of parameters. Furthermore, it gives good results for most of the western languages without any special tuning (except for German which uses capital letters for all the nouns). Main characteristic of the method is that it provides very good recall (almost all of the real name-entities are proclaimed as name-entities), but slightly lower precision (some of the proclaimed name-entities are not name-entities) which is in practical setting enough – errors and exceptions are handled separately by the list of exceptions. Empirical evaluation of the method (based on 100 randomly selected news articles) showed precision value of 73% and recall value of 96% – recall and precision are standard Information Retrieval evaluation measures measuring the ‘truth’ and the ‘whole truth’ of the result set.

In addition to the name-entity extraction, we also use name consolidation mechanism which tries to unify different surface forms into one name-entity (e.g. ‘Bill Clinton’==’President Clinton’==’Clinton’).

6 VISUAL INTERFACE

In this section we are presenting the client part of the “Contexter” system offering graphical user interface to the pre-calculated name-entities and bag-of-words representations of the news documents corpus which are stored together with the original textual representation within the database.

The core part of the system is the main graphical user interface form which primarily offers two functionalities:

1. Browsing through the network of connected name-entities (two name-entities are connected if they appear in at least one common document).
2. Visualizing a context of a name-entity appearance within the corpus. The context of a name-entity is shown in three different ways:
 - by a set of keywords usually collocated with the selected name-entity,

- by a set of other name-entities usually collocated with the selected name-entity,
- by a set of keywords collocated with the simultaneous appearance of the selected and most frequent other name-entities.

Usage of “Contexter” consists from the following steps:

1. Preprocessing of the document corpus which generates name-entity and bag-of-words representations which are saved together with the original textual representation within a database. This step is performed only once per database change. Since all the algorithms used in the preprocessing phase are computationally efficient, this step takes approx. 15 seconds for the whole ACM TechNews corpus (11.000 articles) on the 2.4GHz PC. We also experimented with other larger corpora (non English languages) and the experiments showed the system scales linearly according to the size of the data (which is expected according to the design of the system).
2. The user runs the client (from the Figure 1) with the graphical user interface. First the user connects to the database with the three document representations. This loads part of the data into the system (list of all name-entities and cached part of the bag-of-words sparse vectors).
3. The user selects a name-entity from the left most window instantly shows the corresponding content within other three “context windows”. First to the right is the window (1) with the graphical representation of the local context of the network around selected name entity, (2) next, window to the right shows the context from characteristic keywords from the documents where the selected name entity appears, and (3) the right most window shows the context in the form of the most frequent other name-entities collocated with the selected name-entity.

In the next steps the user is selecting other name-entities (either from the complete list on the left, from the graphical interface in the middle or from the right most context list) which instantly adapt the screen according to the new selection. With additional local menu function the user can view the actual context of the documents where the selected name entities appear.

7 DISCUSSION

In this section we are presenting the client part of the In the paper we presented architecture and implementation of the system “Contexter” for analytical browsing of news articles. In the first stage documents from the corpus are preprocessed and transformed into two alternative representations – each document gets in addition to its original textual representation also name-entity and bag-of-words representations. For both transformations we decided to use simple and computationally efficient procedures which give satisfactory results in terms of quality. Quality

could have been slightly increased with the selection of some other methods but on the cost of computational efficiency which would further decrease usability of the interface.

There are several potential additions which might be interesting for the future development of the system. In particular, with a more detailed analysis of the text in the preprocessing stage with natural-language-processing tools we would be able to identify finer grained contexts in which an individual name-entity is appearing; furthermore it would be possible to detect more explicit relationships between name entities. Next, some more text visualization and text summarization techniques can be applied to extend levels of abstraction when observing the contents of the system. With better name-entity recognition and consolidation (disambiguation) the usability of the system would increase – in some case the cost of the preprocessing phase (in terms of time and human resources) is not of such importance. Finally, the whole system would benefit a lot from a wider Human-Computer-Interaction study which would evaluate current system and suggest corrections to the user interface design and to the needs for various user profiles. In the current stage we designed system mainly for research journalists from some of the Slovenian daily newspapers which contributed suggestions through descriptions of their needs and what they perform in their everyday routine.

References

[1] Geroimenko, V., Chen, C. (ed): Visualizing the Semantic Web. Springer Verlag, (2003).

[2] Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999).

[3] Fayyad, U., Grinstein, G., Wierse, A.: Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann (2001).

[4] Chakrabarti, S.: Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufman (2002).

[5] Jurafsky, D., Martin. J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall (2000).

[6] Chen, C.: Visualization of Knowledge Structures, In “Handbook of Software Engineering and Knowledge Engineering”, World Scientific Publishing (2002).

[7] Chase, P., D’Amore, R., Gershon, N., Holland, R., Hyland, R., Mani, I., Maybury, M., Merlino, A., Rayson J.: Semantic Visualization. ACL-COLING Workshop on Content Visualization and Intermedia Representation.

[8] Content Visualization and Intermedia Representations (CVIR’98), <http://acl.ldc.upenn.edu/W/W98/>

[9] Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Comp. Graphics, V8, No.1, 2002.

[10] Pacific Northwest National Laboratory, Information Visualization, <http://www.pnl.gov/infviz/>

[11] Grobelnik, M., Mladenic, D.: Efficient visualization of large text corpora. 7th TELRI, Info. in Corpora (2002).

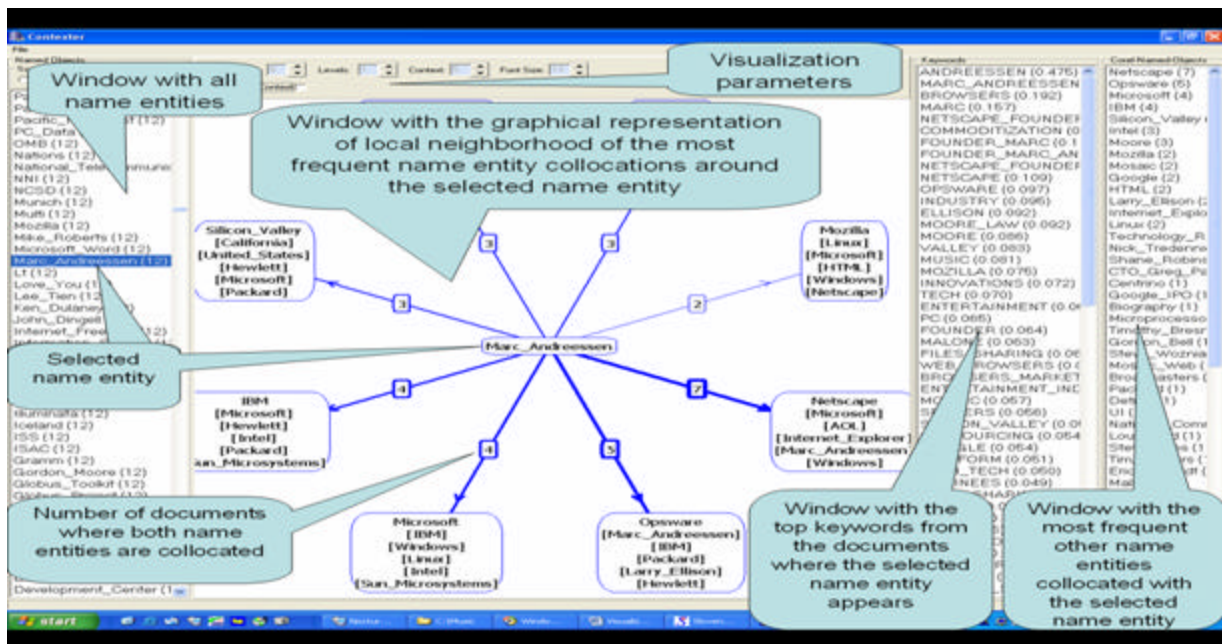


Figure 1. Graphical interface of “Contexter” for browsing/visualizing the name-entity network.