EXTRACONN: Extraction and Analysis of Company Networks from News

Markus Döhring

SAP Research CEC Darmstadt, Bleichstraße 8, Darmstadt, Germany
markus.doehring@sap.com

Abstract. Companies in nowaday's business world have started to recognize the value of Social Network Analysis not only for the discovery of interhuman relations, but also for the examination of e.g. interorganizational dependencies and collaboration patterns. However, the corresponding IT-supported analyses require a large amount of explicitly structured data, which is rarely available and costly to create manually. The contribution of our presented approach to the above described problem is therefore two-fold. The first challenge is to bridge the gap between unstructered text data and the required network structures for the specific domain of business news. The aim is to obtain a network of companies that represents relevant facts (like industry relatedness) as close to reality as possible. We will then apply a chosen catalogue of social network analysis techniques in order to validate their suitability and to generate immediate feedback for the network generation step.

1 Background

1.1 Motivation and Scope

The application of concepts from research in the field of Social Network Analysis (SNA, [1] and [2]) has just recently gained ground in the business domain. The focus thereby is not solely on the relations between individuals, but also on the relations between organizations. Companies begin to recognize the value of locating experts and leads, identifying patterns in information and communication flows, supporting industry ecosystem analysis through uncovering critical linkages between companies, and much more, all supported by SNA concepts and tools [3]. However, one drawback which might prevent the extensive exploitation of IT-supported SNA techniques is the large effort needed for the creation of the required data structures. If the desired networks are not already available in a structured and computer-processable format (which is rarely the case), they usually have to be manually created by domain experts.

Contrasting the prevalent lack of prestructured networks, business organizations usually dispose of large sets of unstructured data, which may be publicly available or kept private within the organization. This data often contains very valuable information, which is potentially exploitable also for SNA. Particularly the business field may benefit from leveraging on this data, since software for

SNA is already available and can be employed provided that the relevant relations are properly extracted from the unstructured information and prepared according to SNA requirements.

We therefore address the (semi-)automatic extraction of business entities, which in our case correspond to companies, as well as relations together with the subsequent creation of network structures and the application, validation and eventually extension of SNA techniques in a holistic manner. To delimit the examined domain also with respect to the analyzed data, we will focus on the specific domain of business news articles. EXTRACONN (**Extr**action and **A**nalysis of **Co**mpany **N**etworks from **N**ews) within THESEUS/TEXO[1] is a small research project conducted by two students in cooperation with SAP Research within a Master's studies module in Information Science at Hochschule Darmstadt. It is defined for a period of six months and can be extended to twelve months if results seem promising. According to the basic conditions described above, the scope of our intended research approach is two-fold:

1. We need to bridge the gap between unstructured text data and the required network structures. For our domain of business news articles, we aim at obtaining a business network of entities (companies) whose relation is defined by the strength of e.g. their cooperation. One hypothesis we will thereby further examine is that if single text segments can be assigned to corresponding companies, entity co-occurences in text reflect the real-world intensity of business cooperation. Additionally, we hypothesize that available metadata (e.g. company profiles) helps to improve the quality of the resulting business network. This concerns, for example, the achievement of a finer network granularity (i.e. a more meaningful topology) or the filtering of noise and network parts irrelevant for analysis.
2. Having the network structures available, we will evaluate how well standard SNA metrics[2] and procedures help to bring about valuable information. The feedback received from exploratory Social Network Analysis will eventually help to refine the requirements for the network structure generation. The SNA application may include rather intuitive findings, like the determination of key players within the network, or the industrial classification of organizations according to the network topology. As an illustrative example, a company might be interested in the other companies residing in the same "industrial activity cluster" derived from a combination of its internally and externally available unstructured data. It might turn out, that a direct competitor plays a central role in a cluster different from the company's one. A further manual investigation could lead to the conclusion, that the competitor has found sales markets previously unexplored by the first company.
   However, there are also elaborate SNA techniques which can produce results that that are not obviously connected to our use-case (like triad analysis,

[1] see http://theseus-programm.de/scenarios/de/texo

[2] A simple example for a SNA metric would be the calculation of betweenness centrality for a single node in terms of how many shortest paths pass through the regarded node.

which examines the spectrum of how each node within a set of three elements is related to each other), but may nonetheless yield valuable findings, so we will not restrict ourselves with respect to the employed analysis techniques.

1.2 Relation to THESEUS/TEXO

The TEXO use case within the THESEUS-program aims at providing a platform and methodology, which increases the flexibility of future business value networks based on services. Besides the savings in time and therefore financial effort that could be achieved by our targeted semi-automated network generation and analysis process[3], our approach may moreover be used to address the collaboration aspect of business entities interacting within TEXO. In this respect, the results of our work could be a contribution to a recommender system which proposes suitable collaboration partners. For example, if a service consumer is searching for a service that not only fulfills functional requirements but also requirements like "*the service should not be offered by service providers which have critical linkages to my customers*", this will eventually lead to a higher satisfaction of TEXO participants and therefore an increased usage of the TEXO platform. Another benefit could be for the TEXO platform administrator to be able to observe and analyze the emerging business communities of practice, in a sense that some findings gained from the business news domain could be transferred e.g. to the analysis of business expert forums. Informal business networks within TEXO will presumably consist of a number of subnets, which are kept together by only a few key players. If those key players leave the TEXO ecosystem, this may lead to a collapse of the whole subnet. Being aware of such communities of practice means that the TEXO platform administrator will eventually be able to influence those communities, to take preventive actions and to actively foster service-based cooperation.

2 Description of the Approach

We start with the collection and tailoring of a news corpus, where the business entities are ideally already annotated or the set of entities contained within the corpus is given. It would also be desirable to have a business network in an explicitly structured form (suitable for the news dataset) as a gold standard to test it against the retrieved one. For the basic preprocessing (like tokenization and stemming, if necessary) we will rely on standard tools and concepts available from Natural Language Processing (NLP). NLP itself is therefore not within the main focus of our research efforts. For relation extraction, several approaches concerning for instance the extraction based on statistical means (e.g. taking into account word/entity co-occurences) or based on the syntactical level (e.g. sentence structure patterns, see [4])[4] will have to be considered. The syntactical

[3] In contrast to a manual creation and analysis of the networks by domain experts

[4] For example: if a noun phrase is followed by the term "such as", followed by one or several noun phrases, subclass relationships can be derived correspondingly.

level seems to be especially relevant for the finding of ”typed” relations in the network (e.g. competition or collaboration), since statistical concepts are often not sufficient to go beyond the determination of simple relatedness of terms. For working on the syntactical level, more sophisticated techniques from NLP like part-of-speech tagging for sentence structure determination may be borrowed.

For the application of SNA techniques, we intend to make use of the rich concepts elaborately described in [1], [2], [5], and [6]. We will also use existing state-of-the art algorithms (partly available as tools and libraries) and if required tailor or extend them. The general aim is to find and adapt a combination of relation extraction and SNA algorithms that works best on our business news domain.

Finally, the results will have to be evaluated. One way to achieve this would be to visualize the networks together with the findings and to have them validated manually by (business) experts. One could compare, for example, whether discovered industry clusters based on the networks intuitively make sense. Furthermore, having the ability to browse the network in an exploratory way offers the opportunity to discover potentially valuable coherences formerly not covered by SNA. Validation could also be performed automatically. Regarding the described example of determining key players, it could be checked whether they correspond to market leaders measured by market share, given that corresponding lists are publicly available in computer-readable form.

Figure 1 summarizes the described approach from data collection to network creation and analysis, including feedback generation. We are currently processing the first step (data collection and corpus generation). However, we intend to finish a first iteration of the approach until the end of April 2008. This will only include the implementation of straight-forward and existing term-occurrence based algorithms for network generation together with a plain visualization. We expect to take the results as indicators for further improvements of our approach and starting points for deeper research. Until August this year, we aim at having a clear picture of the concepts required for proper network generation and of how to apply suitable SNA techniques, together with an estimation of the actual (business) value of the analysis results. This can be used as a go or no-go indicator for further research proceedings, which could last until February 2009.

3 Related Work

Our approach implies investigations in fields of research adjacent to or involved in information extraction [7] in order to see which concepts could be transferred and employed. Obvious general adjacent areas are ontology learning [8] and data mining [9]. Available tools can be tested and eventually be built upon.

The exploitation of news datasets for research purposes is widely common due to the high availability, their large overall text volume and a certain degree of inherent structure (e.g. according to certain categories). Nonetheless, little work exists that has already made use of news in a similar context and to a similar extent as we intend to.

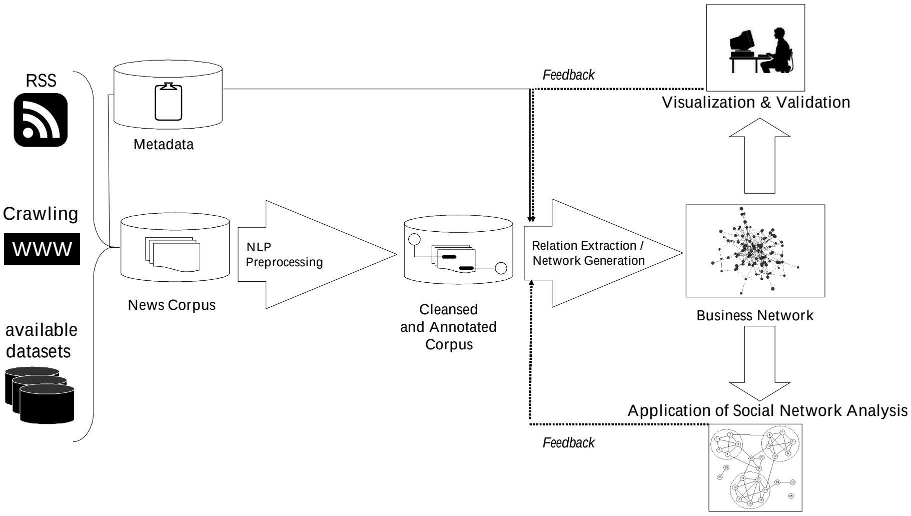
RSS
Crawling
WWW
available datasets
Metadata
News Corpus
NLP Preprocessing
Cleansed and Annotated Corpus
Relation Extraction / Network Generation
Business Network
Visualization & Validation
Application of Social Network Analysis
Feedback
Feedback

Fig. 1. Summary of our Approach to Company Network Generation and Analysis

Gründel et al. [10] address the fields "information extraction", "subject classification" and "emotional classification" in financial news via statistical means, like Hidden Marcov Models or Support Vector Machines. Laegreid and Sandal [11] also apply information extraction techniques to discover coherences in news texts, but rather focus on the performance and tuning of different Part-of-Speech and Named Entity Recognition taggers, while the visualization and analysis of business networks is only outlined in the "further work" section. Jin et al. [12] describe concepts for entity and relation extraction from unstructured text (e.g. for determining characteristics and strengths of entity connections based on web search). For example, the names of two companies and "lawsuit" are issued as search terms and the retrieved documents are further analyzed for relation generation. However, they do not dive deeper into a subsequent topological network analysis. Bernstein et. al. [13] take an approach similar to ours, although the methods used for network generation are exclusively occurrence-based and the SNA techniques they apply are restricted to the determination of key players and industry clusters.

References

1. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press, Cambridge (1994)
2. Scott, J.P.: Social Network Analysis: A Handbook. SAGE Publications (January 2000)
3. Ehrlich, K., Carboni, I.: Inside social network analysis. Technical report, IBM Watson Research Center (2005)
4. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. Technical Report S2K-92-09 (1992)

5. Hanneman, A., Riddle, M.: Introduction to social network methods. (2005)
6. de Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York, NY, USA (2004)
7. Appelt, D., Israel, D.: Introduction to information extraction technology: A tutorial prepared for ijcai-99. SRI International (1999)
8. Maedche, A., Staab, S.: Ontology learning from text. In: NLDB ’00: Proceedings of the 5th International Conference on Applications of Natural Language to Information Systems-Revised Papers, London, UK, Springer-Verlag (2001) 364
9. Maimon, O., Rokach, L., eds.: The Data Mining and Knowledge Discovery Handbook. Springer (2005)
10. Gründel, H., Naphtali, T., Wiech, C., Gluba, J.M., Rohdenburg, M., Scheffer, T., Dabiri, G.: Clipping and analyzing news using machine learning techniques. In: Proceedings of the International Conference on Discovery Science. (2001)
11. Lgreid, T., Sandal, P.C.: Financial news mining: Extracting useful information from continuous streams of text. Master’s thesis, Norwegian University of Science and Technology, Faculty of Information Technology, Mathematics and Electrical Engineering, Department of Computer and Information Science (2006)
12. Jin, Y., Matsuo, Y., Ishizuka, M.: Extracting a social network among entities by web mining. In: Proc. of ISWC2006 Workshop on Web Content Mining with Human Language Technology. (2006)
13. Bernstein, A., Clearwater, S., Hill, S., Perlich, C., Provost, F.: Discovering knowledge from relational data extracted from business news. In: Proceedings of the Workshop on Multi-Relational Data Mining at KDD-2002, University of Alberta, Edmonton, Canada (2002) pp. 7–20

Disclaimer

The TEXO project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference ’01MQ07012’. The authors take the responsibility for the contents. The information in this document is proprietary to the following Theseus Texo consortium member: SAP AG. The information in this document is provided ’as is’, and no guarantee or warranty is given that the information is fit for any particular purpose. The above referenced consortium member shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law. Copyright 2008 by SAP AG.