Causal Relation Detection for Activities from Heterogeneous Sources

Philipp Katz and Alexander Schill

Technische Universität Dresden, Germany
Department of Computer Science, Chair of Computer Networks
philipp.katz@tu-dresden.de

Abstract. On the web, information representing specific activities is often scattered over different systems. Although, causal relations exist between these activities, these are usually not obviously visible to the user, unless explicitly given. This paper outlines the difficulties which are caused by missing relations. The core contribution of this work will be a system which is capable of identifying cause-effect relations between single activities. The system will use these relations to form coarse-grained groups consisting of sequences with single activities. The intended goal is to employ the detected relations to reduce information overload while increasing accountability, clarity, and traceability for its users. The research is conceived under the assumption of handling heterogeneous sources of information. A further objective is to create a highly generic and flexible system which can be adapted to different use cases. The system will be evaluated with concrete case studies, one of them analyzing relations on software development sites such as SourceForge.

Keywords: Internet Information Extraction, Information Aggregation, Information Integration, Relation Extraction, Data Linking.

1 Introduction

Over the last years, the WWW has evolved into a highly dynamic and interactive medium. In conjunction with the buzzword "Web 2.0", so called user-generated content, published on different platforms such as blogs, wikis, social networks, or media portals, is gaining influence. As well as the number of different sources, the amount and frequency of generated information is increasing continuously. Besides the general and often discussed phenomenon of "information overload", which has already been characterized in various different sources [1], a further problem can be observed: "information scattering", where information concerning one specific topic is usually published across various different sources. In [2], information scattering is described as a situation, where few sources exist, "that contain many items of relevant information, while most sources have only a few". Although, their analysis is influenced by a standpoint of library science, their general notion can be transferred to the heterogeneity and diversity of the WWW, including user-generated content.

A. Harth and N. Koch (Eds.): ICWE 2011 Workshops, LNCS 7059, pp. 312–316, 2011.

2 Terminology and Scenario

In this work, the term "activity" is defined as an atomic event occurring at a certain time. A source propagating activities is called "activity generator". Obviously, activities exist which are triggered by other activities, leading to causal relations between pairs of activities, which can be considered as "cause-effect relations".

To substantiate the idea behind this work, consider the following concrete usage example: The workflow of a typical open source software project is organized using various tools such as an issue tracker, a version control system, discussion forums, and mailing lists as depicted in Fig. 1. End users experiencing issues with the software, use the forums or mailing lists to start a discussion. Different users and developers are involved in this discussion, and finally a bug report is posted to the tracker. After a period of time, a developer reads about the problem, commits a bug fix to the version control system, and closes the ticket. The described scenario consists of various activities; the bug report is triggered by the forum discussion, the committed fix is triggered by the bug report.

The scenario outlines the difficulty to get the current situation of the project. While there obviously exists a latent sequence of activities triggered by the initial discussion, it is later difficult to reconstruct such a process, as events are scattered over different heterogeneous sources. Decisions which had an impact on activities are hard to trace from a retrospective point of view. Support is complex, as users need to search for information concerning a specific problem on different sources, manually synthesizing a causal chain representing a decision process.

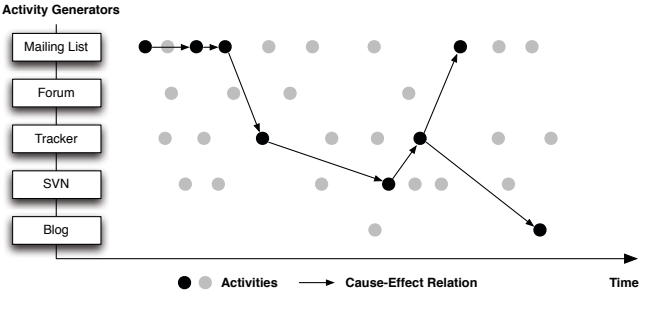

Fig. 1. Example for an activity flow in an open source software project

3 Research Questions

Based upon the outlined scenario in the preceding section, the following three research questions form the main contributions of this work:

How can a system to reduce information overflow and scattering be designed? The goal of this work is to address the problem of information overflow by establishing relations between single pieces of information, forming semantically related groups. Such a group can be characterized as an abstraction of single, fine-grained pieces of information, which are connected by cause-effect relations. Such representations help users to gain a view on "the global picture", making it possible to understand compound, scattered activities formed of single events. The notion of activities will be employed to reduce the information overflow, allowing the user to filter out irrelevant information, only getting into detail where applicable. To establish these relations in a precise way, suitable algorithms need to be employed and necessary components of the system need to be identified.

How can the system cope with highly heterogeneous types of data? The sources considered by the system can be characterized as highly heterogeneous. Where, from a low-level and technical standpoint, the term "heterogeneous" can be used to describe varying formats and standards, on a more abstract level, a great spectrum of quality in respect to the actual data has to be considered. Different sources usually provide diverse amounts and types of metadata. Besides explicit features which can be extracted directly from their sources, also implicit information will be taken into account, which needs to be induced by the system automatically. This might include correlations between different authors, temporal aspects, or characteristics concerning different sources. Those features will be employed to create a generic model for the data, which forms the foundation for the applied algorithms. Due to the heterogeneous nature of the considered data, the potential sparsities and unequal distributions concerning the presence of those features need to be taken into account.

Which methodologies can be applied to ensure the adaptiveness of the system to various usage scenarios and domains? The general aim of this work is a highly generic and flexible system, which is capable of covering a wide range of potential use cases. Therefore, a parametrizable and configurable framework which can furthermore be integrated into various existing workflows needs to be provided.

4 Related Work

In the domain of network monitoring and management, event correlation and filtering systems (ECS) are employed to diagnose network failures. Various proposals for approaches and practical implementations exist [3,4,5], sharing the general goal to identify "root causes" for specific problems and to filter the massive amount of single events by correlating and aggregating them to more abstract, "conceptual" events [6]. Obviously, these approaches focus on technical events generated by machines such as servers or routers. To the author's knowledge, there have been no efforts on mapping the concepts from ECS to honor the specific properties of event streams with content generated by and for human users.

Research in the area of Topic Detection and Tracking (TDT), which was initiated under a DARPA program, is analyzing textual and audio-visual data from news broadcast sources in order to perform a topic based characterization and to detect links between pieces of information [7]. More recently, efforts have been conducted to identify events in data from community and social media sources such as Flickr, Youtube, or Facebook [8,9,10] employing document clustering techniques to identify groups of information associated to a specific event. As a general constraint, each instantiation of the mentioned approaches allows to consider only one specific source, disregarding the possibility to establish relations between different types of heterogeneous data spread over various sources.

Work on Process Mining deals with extracting process models from event log data [11]. However, this data is generated by technical systems and therefore conforms to a very strict, well-defined and homogeneous structure which is in general not reflected by user-generated content on which this work puts its focus.

5 Current Progress

For an initial analysis, a crawler for SourceForge[1] was developed. The crawler was used with the phpMyAdmin project, aggregating activity feeds, tracker data, messages from mailing lists, forum posts and commit messages from their source code repositories. The resulting dataset consists of over 180,000 individual items spanning a time interval of approximately ten years. After a first naïve experiment using text clustering techniques based on a plain term model to identify causal relations, it can be concluded, that more elaborate approaches need to be employed in order to achieve reasonable results.

The outlined dataset contains a set of explicit features which can be taken into consideration for creating connections between pairs of items. These features include hyperlinks between items or indicators such as revision or tracker IDs given in the text, which can be extracted using patterns. They will be employed for building a preliminary baseline and for evaluating the initial results. Further iterations will be measured with regard to this baseline.

With PRISMA[2], a system architecture for handling the information overload in an enterprise context has been described [12]. The implementation of this system will provide the framework for evaluating the algorithms from this work.

6 Future Work Plan

In the short term, the future work will focus on extracting further features from the given dataset. Therefore, an extensive feature engineering will be performed, evaluating implicit features which can be extracted from the data. In general, these might include structural, temporal, statistical, linguistical or semantic features, regarding individual properties.

[1] `http://sourceforge.net/`

[2] PeRsonalization of Information StreaM Aggregates.

In the medium term, following a bottom-up approach, an abstraction from the knowledge gained from practical experiments concerning the data from SourceForge will be performed. Regarding further use cases, an evaluation is neccessary, on how the model needs to be generalized and extended, in order to allow for the aimed adaptiveness. Therefore, a further concrete scenario which considers the domain of Wikipedia content will be described. The objective is to research the impact of pieces of news published by different sources on Wikipedia articles. Evidently, a great amount of edits performed in the Wikipedia is triggered by current events of the day. In the long term, an instantiation of the system will be used as a component within PRISMA.

Acknowledgments. The PRISMA project is funded by the Free State of Saxony and the EU (European Social Fund).

References

1. Richtel, M.: Lost in E-Mail, Tech Firms Face Self-Made Beast, `http://www.nytimes.com/2008/06/14/technology/14email.html` (retrieved March 7, 2011)
2. Bhavnani, S.K., Wilson, C.S.: Information Scattering. In: Encyclopedia of Library and Information Sciences, 3rd edn., pp. 2564–2569 (2010)
3. Vaarandi, R.: Platform Independent Event Correlation Tool for Network Management. In: Proc. of the 2002 IEEE/IFIP Network Operations and Management Symposium (2002)
4. Liu, G., Mok, A.K., Yang, E.J.: Composite Events for Network Event Correlation. In: Proc. of the Sixth IFIP/IEEE International Symposium on Integrated Network Management, pp. 247–260 (1999)
5. Jiang, G., Cybenko, G.: Temporal and Spatial Distributed Event Correlation for Network Security. In: Proc. of the 2004 American Control Conference, vol. 2, pp. 996–1001 (2004)
6. Hasan, M., Sugla, B., Viswanathan, R.: A Conceptual Framework for Network Management Event Correlation and Filtering Systems. In: Proc. of the Sixth IFIP/IEEE International Symposium on Integrated Network Management, pp. 233–246 (1999)
7. Allan, J.: Introduction to Topic Detection and Tracking. In: Topic Detection and Tracking: Event-Based Information Organization, Springer, pp. 1–16. Springer, Heidelberg (2002)
8. Becker, H., Naaman, M., Gravano, L.: Learning Similarity Metrics for Event Identification in Social Media. In: Proc. of the Third ACM International Conference on Web Search and Data Mining, pp. 291–300 (2010)
9. Sayyadi, H., Hurst, M., Maykov, A.: Event Detection and Tracking in Social Streams. In: Proc. of International Conference on Weblogs and Social Media, ICWSM (2009)
10. Zhao, Q., Mitra, P., Chen, B.: Temporal and Information Flow Based Event Detection From Social Text Streams. In: Proc. of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1501–1506 (2007)
11. van der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Information Systems 36(2), 450–475 (2011)
12. Katz, P., Lunze, T., Feldmann, M., Röhrborn, D., Schill, A.: System Architecture for handling the Information Overload in Enterprise Information Aggregation Systems. In: Proc. of the 14th International Conference on Business Information Systems (2011)