Succinct summaries of narrative events using social networks

Bart de Goede, Maarten Marx, Arjan Nusselder, Justin van Wees
ISLA, University of Amsterdam
Science Park 904 1098 XH Amsterdam, The Netherlands

ABSTRACT

This paper addresses the following research aim: provide a useful but succinct summary of long narrative events involving the interaction of several speakers. The summary should enable users to navigate to specific parts of the event using hyperlinks.

Our solution is based on a representation of the main actors of the event and their interactions as a social network. The solution is applicable to events in which these interactions are more or less formally structured and detectable. This includes theatre and radio plays, recordings of a scientific workshop, proceedings of parliament and meetings notes in general.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

1. INTRODUCTION

This paper describes how we used social network techniques [24] to provide hyperlinked summaries of long documents describing debates in parliament. We generalize from this specific use-case and show how the same technique is applicable in a variety of cases, with examples from the work of Shakespeare. We describe the technical requirements on the input data and provide the technical details of creating the hyperlinked summaries. Except for minor input preprocessing and the generation of pictures all processing can be done using XML technology, in particular XSLT.

Input suitable for our approach has a narrative structure, a clean linear order, several actors or speakers providing information, and, most importantly, the interaction patterns between the actors indicate interesting parts of the data. Of course, there should be a need for summarizing the document. This can be because it is either long (like a conference proceedings), in a non-intended format (like the written version of a theatre play), or in a format which is difficult to browse (an audio or video document).

Typically the data is a representation of an event and available in different media formats. Here is an example: Imagine you participate in a workshop with many lively discussions during the talks. The formal proceedings with papers is not a good reflection of the actual going-ons. Rather one would like to have a full video and audio recording of the talks and discussions, preferably aligned with the slides, and even better with a written verbatim record. The latter can be used for search.

The purpose of our summaries of these rather inaccessible documents is twofold [19]. First, divide the documents into natural parts which can be hyperlinked to and which are good entry points to the documents. Secondly, the summary should indicate the potentially interesting parts of the document. It is instructive to compare this to an information retrieval setting. The first purpose then is to define the units of retrieval. The second is to create a prior probability for ranking search results, even in the absence of a query [18]. The link structure of documents (e.g PageRank) is an important indicator of this "interestingness prior" [16].

Using social interactions as link structure is the idea behind our summaries. The work of Koolen and Kamps has shown that both local and global link-structure can successfully be applied to calculate interestingness of documents or parts of documents. In fact, often the simplest baseline — the more in- and outlinks, the more interesting— is hard to improve [16].

The problem with many vertical search engines is the absence of hyperlinks in their documents. Without using other (specialized) techniques, this makes their relevance ranking only dependent on the textual content of the documents. This often results in a search experience which is inferior to what people are used to with Google [12].

Following the pioneering work of `http://theyworkforyou.com` in 2004, several vertical search systems for parliamentary proceedings have been created, often by developers not linked to the government. The Wikipedia page `http://en.wikipedia.org/wiki/Parliamentary_informatics` contains a somewhat complete list of these sites. A large problem for these systems is the absence of hyperlinks in the documents. Turning the corpus of documents into a hyperlinked network is seen as a key to success, but also as the most difficult and costly part of the system. The most important hyperlink is from Named Entities in the text to a representation (e.g., a biographical page) of these politicians.

This is not trivial as it involves all difficulties of data deduplication, in particular on scanned and OCRed material [7] The advantage of these hyperlinks together with a good seg-

ACM Hypertext Eindhoven

mentation of the text is that one may compute views on the corpus very much like database views [10]. The most obvious is a list of all activities of one member of parliament with hyperlinks going into the proceedings.

The second advantage is that the segmentation of the input text into meaningful parts together with the normalization of named entities caused by the addition of hyperlinks enables a rich search experience. Two important techniques that become available are faceted search and entry point retrieval.

Faceted search [11, 12] is a powerful technique for combining free keyword search with additional restrictions on metadata attributes. Good examples are eBay and LinkedIn. To assign meaningful facets to search results, the unit of retrieval has to be uniform and somehow natural. If the unit of retrieval is the complete proceedings of one day in parliament (as customary in most search systems employed by parliaments today), basically only the date makes sense. If it is smaller then semantically interesting facets become meaningful: the topic of the debate with values from a controlled vocabulary, or, when the unit of retrieval is individual speeches, the name, party and function of the speaker.

The search engine `http://polidocs.nl` shows for each query the top five years, politicians and parties with most hits on that query. This feature of the search engine was rated most useful by expert users, in line with user studies reported in the literature [12]. This simple device has two powerful functions: it gives the user quick insights in the distribution of the search results on a number of important dimensions (when, who). It also provides a simple way to restrict the query to a certain facet: simply clicking on the facet value restricts the returned hits to those for which the value holds.

Best entry point retrieval [22, 8] refers to an alternative to the well known Google style search in which documents are returned as units of retrieval. In contrast, entry point retrieval provides for each relevant document a small number of entry points in the document which best fit the information need expressed in the query. Such an entry point retrieval system has the following technical prerequisites: documents can be automatically partitioned into meaningful units; each unit can be given a unique name, so that it can be hyperlinked; and documents are in such a format that devices like browsers can bring the user directly to these entry points.

2. VISUAL SUMMARIES USING STRUCTURE

We applied the techniques discussed above in a visual summarization tool for meeting notes. The tool consists of three parts. Given the proceedings of a meeting, first a social network of the speakers of the meeting is created in which arrows indicate interruptions (Figure 1). This gives a high level overview and indicates “where the action took place”. Each node is hyperlinked to a structural summary of the speech of that person including all interruptions. Figure 2 contains these timelines for six of the nodes in Figure 1. Each timeline partitions the speech into parts spoken by the speaker (indicated by the red vertical mouths) and those spoken by others (the blue mouths). The size of each mouth is a measure for the number of words. The first interruption by a person is labeled by his name. Thus the second figure presents a fine structure of the incoming arrows (the indegree) of a node, and preserves the temporal order.

Each mouth is hyperlinked to the specific part in the debate. This hyperlink brings the user to the third part, the concrete notes of the debate.

Theatre plays. We have successfully applied this technique to meeting notes of several parliaments but also to less formal meeting notes like sessions of the student council of our faculty [6]. To show that it is even more universally applicable we also applied it in a different domain: theatre plays.

Also theatre plays have enough interaction between the players to create useful summary networks. We exemplify this using the work of William Shakespeare. All plays of Shakespeare have been turned into XML by Jon Bosak and are available at `http://www.cafeconleche.org/examples/shakespeare/`. They are all valid with respect to one DTD, also available at that site.

From the plays in XML we created two networks. The first is an undirected one with a definition which is often used in creating “old boys networks”: a bigraph of actors and clubs with membership relations between the actors and the clubs induces an undirected network on the set of actors as follows: two actors are related if they belong to the same club. Within a theatre play this idea can be turned into “two actors are related if they participate in the same scene”. Figure 3 shows this network. We used a weighted representation: the thickness of the links is proportional to the number of scenes in which source and target co-appear.

The second network is closer related to the parliament network. We made a crude approximation of the notion *source speaks to target* by creating an edge from source to target if a speech of target immediately follows a speech of source. As before, the number of times this occurs determines the thickness of the arrow. See Figure 4.

3. TECHNICAL ASPECTS

We briefly describe the techniques used in creating the hyperlinked summaries. Our main constraint was scalability: everything had to be done automatically.

Shakespeare.

As the input was available in XML it was easy to process. With XSLT the relevant parts can be given URI’s to make them hyperlinkable. Network data has a standardized XML format (`http://graphml.graphdrawing.org/`) and thus XSLT is the obvious choice for creating networks.

Parliamentary proceedings.

Here things are not that easy, as proceedings are available in all kinds of formats (Word, PDF, HTML, and, with luck, XML). On the other hand, proceedings have a consistently applied and detectable structure. We have experimented with proceedings from nine parliaments[1] and all could be processed using the following general pipeline:

Step 1 Turn into well formed UTF-8 encoded XML.

Step 2 Recognize structure and named entities and give these unique URI’s.

[1] The Netherlands, Belgium, Germany, Spain, EU parliament, Denmark, Sweden, Norway and the UK.

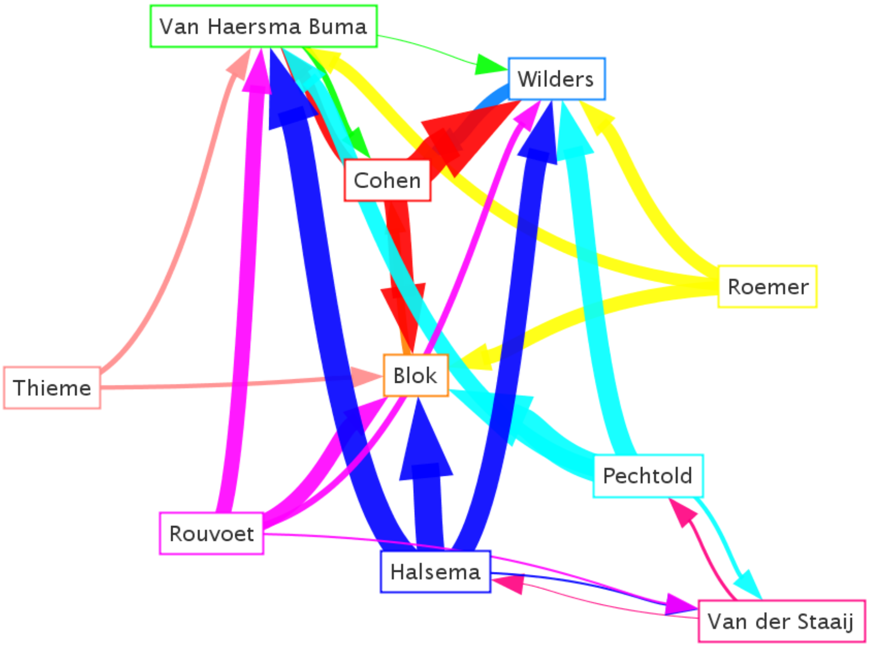

Figure 1: Interruption graph of the plenary debate in the Dutch Parliament of 2010-11-26. An arrow from source to target indicates that source interrupted target during her speech. Thickness indicates number of interruptions. Source of data: `https://zoek.officielebekendmakingen.nl/h-tk-20102011-13-7.html` GraphML file: `http://data.politicalmashup.nl/debates/nl/h-tk-20102011-13-7.1.graphml`

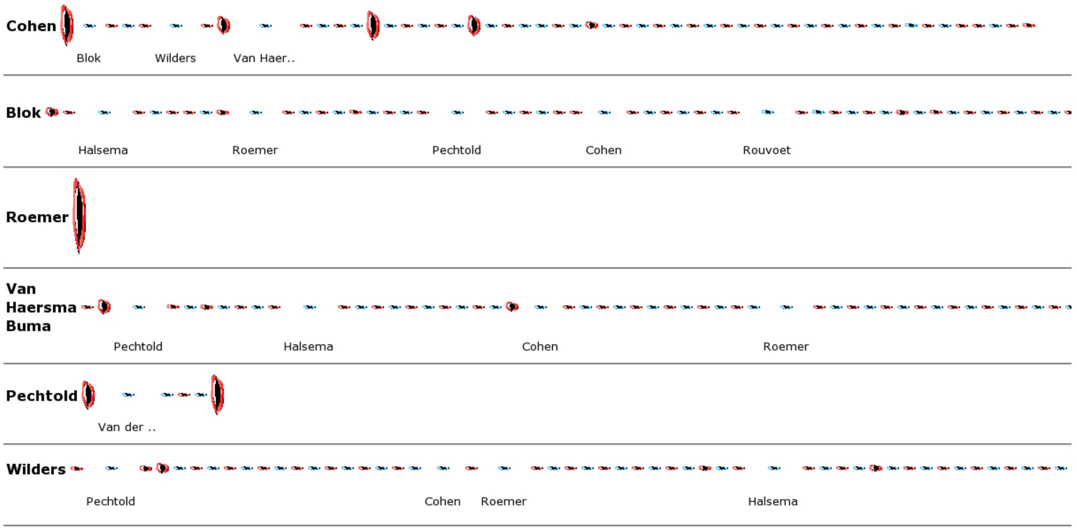

Figure 2: Timeline of the same debate as depicted in Figure 1. Every line of mouths shows the speech of one speaker with all interruptions. In effect it shows the fine-structure of the indegree (the incoming arrows in Figure 1) of the actor. Red mouths refer to speeches of the speaker, blue to interruptions. The size of the mouth is proportional to the number of words spoken. Every mouth contains a hyperlink to that exact part of the debate. Source of data: `https://zoek.officielebekendmakingen.nl/h-tk-20102011-13-7.html` Page on the web: `http://data.politicalmashup.nl/debates/nl/h-tk-20102011-13-7.1-tijdslijn.html`

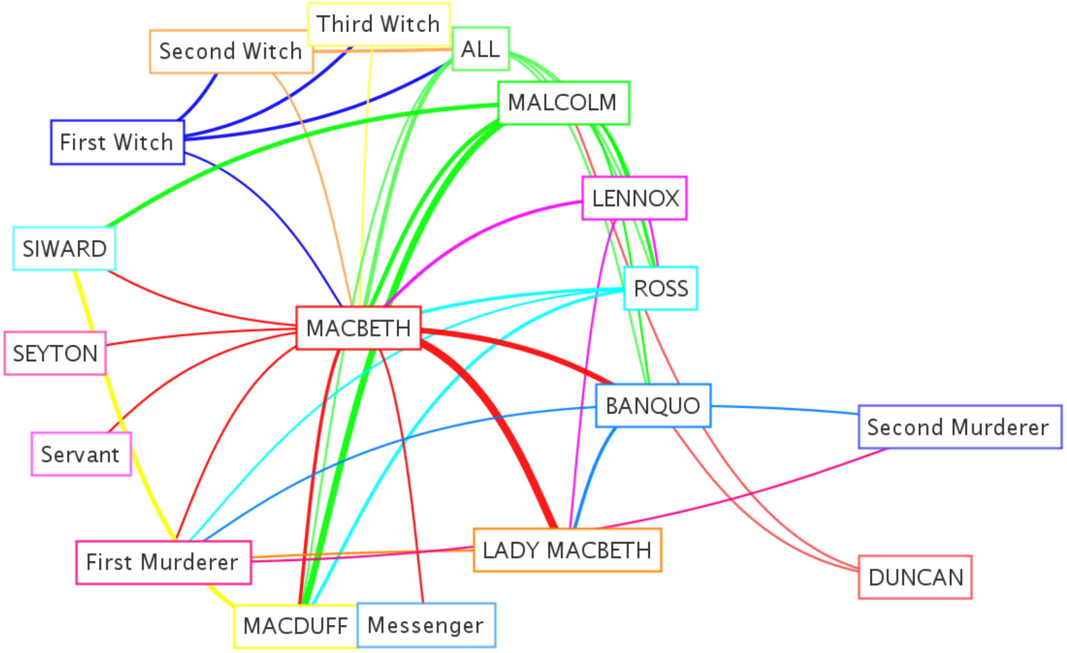

Figure 3: Sharing a scene network of Shakespeare's Macbeth. Only actors are shown which have at least 5 speeches and which are connected by a path to Macbeth. Only edges are shown between actors if they appear in at least two scenes together.

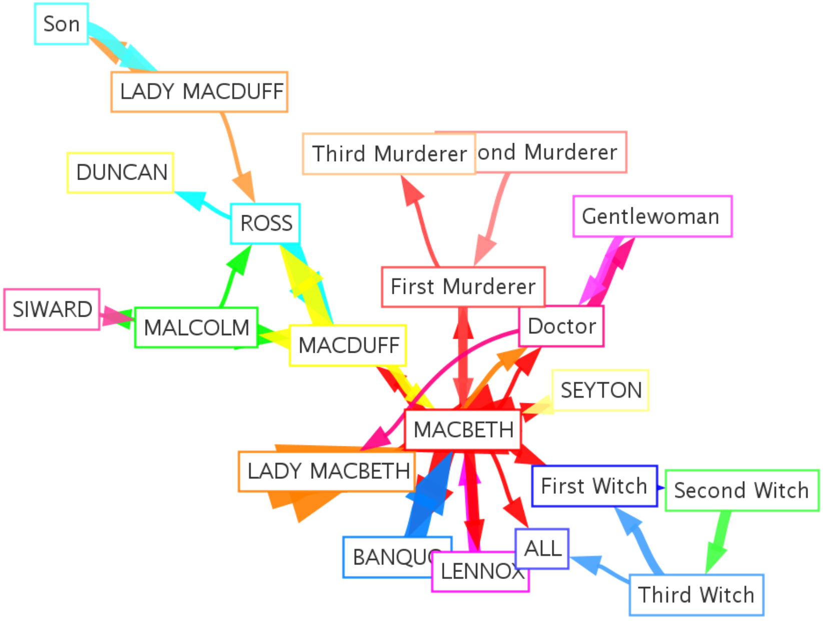

Figure 4: Who speaks to whom network of Shakespeare's Macbeth. Only actors are shown which have at least 5 speeches and which are connected by a path to Macbeth. Only edges are shown if source spoke at least 4 times to target.

Step 3 Parse into a fixed XML format and validate against a Relax NG schema.

Step 4 Create summaries and a hyperlinked corpus.

Step 1 uses a variety of transformation and cleaning techniques. Step 2 uses parliament-specific hand crafted rules, which are expressed declaratively in XPath 2.0. Step 3 is performed using a uniformly applicable XSLT 2.0 transformation. Step 4 uses the technique described above for the Shakespeare plays.

4. EVALUATION

Our system was evaluated in three ways. First, we did a user study with 12 participants of one year of meetings of our faculty student council. Users were predominantly positive and reported they could find back relevant parts easier and faster [6].

In a second evaluation we used the link structure of debates to determine interestingness, similar to the well known PageRank algorithm. Out of a set of 11 features, link structure was the best predictor of interestingness [13].

A third evaluation was done using a panel of expert users of parliamentary proceedings: two journalists from national Dutch newspapers and two civil servants working in the information department of the Dutch parliament. The Dutch proceedings do not have a table of contents and only after each parliamentary year a non-electronic register is created. Users reported that the network presentation greatly facilitated finding the “hot spots” inside the long (and often boring) meeting notes. Being able to drill down on specific persons and parties was also highly appreciated.

5. RELATED WORK

The notion of a semantic-network based discourse was from early on an objective within hypertext research [5, 9]. Since then we have seen further developments on modelling argumentative discourse in general [4], sophisticated requirements for scholarly argumentation [3, 15], or establishing large narratives [23]. Linking plays an important role in these approaches, as the mechanism to represent the dynamic and rhetoric of hypertext, a theme common in hypertext literature [1, 17, 21]. More recently new approaches have emerged that make use of cognitive hyperlinking for the generation of interactive stories [20]. Though these works provide a solid body of work we make use of in our work, there is also criticism about the impact of hypertext technology in the sphere of literature [2]. With our work we respond in particular to Bernstein’s arguments about usability and testing.

6. CONCLUSION

We showed that turning implicitly available link structure in narrative events like meeting notes into explicit hyperlinks can successfully be applied to create high level summaries of otherwise difficult to access documents. Future work may include the use of other summarization techniques such as wordclouds [14] and tight integration with search technology.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 380-52-005 (PoliticalMashup).

7. REFERENCES

[1] Mark Bernstein. Patterns of hypertext. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, HYPERTEXT ’98, pages 21–29, New York, NY, USA, 1998. ACM.

[2] Mark Bernstein. Criticism. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT ’10, pages 235–244, New York, NY, USA, 2010. ACM.

[3] S. Buckingham Shum and A. Selvin. Structuring discourse for collective interpretation. In *Electronic proceedings of Open Conference on Collective Cognition and Memory Practices*, 2000. `http://www.limsi.fr/WkG/PCD2000/indexeng.html`.

[4] Locke M. Carter. Arguments in hypertext: a rhetorical approach. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, HYPERTEXT ’00, pages 85–91, New York, NY, USA, 2000. ACM.

[5] George H. Collier. Thoth-ii: hypertext with explicit semantics. In *Proceedings of the ACM conference on Hypertext*, HYPERTEXT ’87, pages 269–289, New York, NY, USA, 1987. ACM.

[6] G. de Hollander and M. Marx. Summarization of meetings using word clouds. In *under submission*, 2010.

[7] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007.

[8] N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, editors. *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, volume 4862 of *LNCS*. Springer, 2008.

[9] G. Halasz, Frank. Reflections on notecards: seven issues for the next generation of hypermedia systems. *Commun. ACM*, 31:836–852, July 1988.

[10] A. Halevy. Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294, 2001.

[11] M. Hearst. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR Workshop on Faceted Search*, 2006.

[12] M. Hearst. *Search User Interfaces.* Cambridge University Press, 2009.

[13] M. Jongmans. Using IT knowledge to improve knowledge about political-economic space. Master’s thesis, Erasmus University, Rotterdam, The Netherlands, 2010. `http://politicalmashup.nl/uploads/2010/12/thesis-2.pdf`.

[14] R. Kaptein, D. Hiemstra, and J. Kamps. How different are language models andword clouds? In *Proceeings ECIR*, pages 556–568, 2010. urlhttp://dx.doi.org/10.1007/978-3-642-12275-0_48.

[15] David Kolb. Scholarly hypertext: self-represented complexity. In *Proceedings of the eighth ACM conference on Hypertext*, HYPERTEXT ’97, pages 29–37, New York, NY, USA, 1997. ACM.

[16] Marijn Koolen and Jaap Kamps. Searching cultural heritage data: Does structure help expert searchers? In *Proceedings of RIAO 2010: Adaption,*

personalization and fusion of heterogeneous information, 2010.

[17] G. P. Landow. The rhetoric of hypermedia: Some rules for authors. In Paul Delany and George P. Landow, editors, *Hypermedia and Literary Studies*, pages 81–103. Cambridge: The MIT Press, 1994.

[18] Ch. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[19] V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2):80–83, 2008.

[20] Joshua Tanenbaum, Karen Tanenbaum, Magy Seif El-Nasr, and Marek Hatala. Authoring tangible interactive narratives using cognitive hyperlinks. In *Proceedings of the Intelligent Narrative Technologies III Workshop*, INT3 '10, pages 6:1–6:8, New York, NY, USA, 2010. ACM.

[21] Susana Pajares Tosca. A pragmatics of links. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, HYPERTEXT '00, pages 77–84, New York, NY, USA, 2000. ACM.

[22] A. Trotman, S. Geva, and J. Kamps. Report on the sigir 2007 workshop on focused retrieval. *SIGIR Forum*, 41(2):97–103, 2007.

[23] Jill Walker. Piecing together and tearing apart: finding the story in afternoon. In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots: returning to our diverse roots*, HYPERTEXT '99, pages 111–117, New York, NY, USA, 1999. ACM.

[24] S. Wasserman and K. Faust. *Social Network Analysis.* Cambridge University Press, 1994.