Exploring Entity Associations Over Time

Reinanda, R.; Odijk, D.; de Rijke, M.

Publication date
2013

Document Version
Final published version

Published in
SIGIR 2013 Workshop on Time-aware Information Access: #TAIA2013: August 1, 2013

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Download date:22 May 2022
SIGIR 2013 Workshop on Time-aware Information Access (#TAIA2013)

August 1, 2013
ABSTRACT
We address the problem of entity-oriented search in the humanities and social sciences domain. We are particularly interested in retrieving entities related to a query entity and finding associations between these entities over time. Evidence from our target end users suggests that it is more informative to view these associations as dynamic phenomena that evolve over time than as static phenomena. We present work-in-progress on methods to extract these associations and their temporal extent, and discuss a way of presenting them in an exploratory search interface. This interface is intended to help users to discover interesting associations between entities over time.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords
Entity association, exploratory search, semantic search

1. INTRODUCTION
A social historian’s or political scientist’s expertise usually revolves around several historical or political figures, anchored in time. For historians, this anchoring in time is important to construct a narrative around a certain entity, or to investigate the cause and effect of a phenomenon. In a traditional approach, this temporal demarcation is established manually through close reading of the source materials [1]. However, as more sources become available, better ways of navigating and selecting documents from large collections are needed [8].

In semantic search, we aim to return information directly to users instead of a list of documents: with their strong focus on temporally anchored entities, historian and political scientist stand to benefit from this search paradigm. Consider the following scenario: from a stream of 10 years of news articles, a political scientist has to gather the following information about a political figure: his popularity, his associations with other entities, events or issues related to him, as represented in the given stream of news articles. We hypothesize that to satisfy such entity-oriented and time-aware information needs, both entity-related and temporal information should be incorporated at query time in the form of summaries or visualizations. Presenting all this information in a single interface provides a valuable setup for exploratory search [2]. With an exploratory search interface equipped with additional semantic information, a historian can answer time-aware questions and discover interesting events or issues involving the entities of interest to him.

Our work is driven by the following main question: how can we retrieve and present associations between entities within a time period to enable discovery around these associations? This involves three distinct but related aspects: (1) entity associations: finding entities related to a query entity, (2) temporal extent of the associations: finding the temporal distribution of said associations in the document collection, and (3) evidence concerning the associations: ranking and presenting the documents in which these associations occur.

We design an exploratory search interface which considers these three aspects. A sketch of the interface is shown in Figure 1. It has three main components. First, a histogram panel, which consists of a histogram of the frequency of the query entity over time. Second, a related entity panel, showing a ranked list of related entities and their association with the query entity over time. Last, an evidence panel, which shows ranked list of documents where the associations are found.

2. RELATED WORK
Our work is related to previous work on entity associations, relation extraction, and time-aware exploratory search.
Entity Association. Bron et al. [2] explore the task of related entity finding, considered in TREC. Several association models based on document-level co-occurrence are introduced, with additional entity type filtering components added to improve the related entity retrieval. In [10], various ways of extracting related entities for constructing entity networks are explored. They focus on extracting related entities based on sentence level co-occurrence on documents returned from a query entity.

Relation Extraction. In relation extraction, the goal is to identify relationships between a pair of entities. Merhav et al. [7] approach this by clustering pairs of entities using words from the context in which the entities occur. In [5], temporally anchored relation extraction is explored. They determine the time window of extracted relations based on time expressions contained in the document content.

Exploratory Search. Time Explorer [6] is an interface to explore how news evolves over time. Matthews et al. extract temporal expressions from text to allow searching the past and the future. Entities are provided as facets in exploring these news articles. Odijk et al. [9] introduce a coordinated environment for time-aware exploratory search. Their interface is mainly focused on discovering the evolution of word meaning over time on a large collection of historical documents. The interface is meant to help historians discover high-level patterns and periods of interest.

Our entity-centric approach to exploratory search differs from the work listed above in the following ways. Firstly, we score and rank related entities to help discover interesting associations. Secondly, we extract and visualize the temporal extent of the association, albeit in a simple way. Thirdly, we provide ways for query refinement through visual elements, thus allowing quick selection of related document set. Fourth, we provide a co-occurrence context of the associations, giving users a summary of what these associations are about. Lastly, we allow views of the documents in a chronological order so as to help our users follow how events unfold over time. We combine these elements in a single interface.

3. ENTITY ASSOCIATIONS

We extract three types of named entity: person, organization, and location from documents with the Stanford NLP tool [4] and store these named entities as an additional field in the document index. Beside a document index, we also construct entity co-occurrence indexes at the document and sentence level. We use ElasticSearch for document storage and retrieval.

Entity associations. We extract associations with two methods: (1) DF, relying on the document frequency of co-occurrence between two entities, and (2) PHI-COEFFICIENT value, by computing the chi-square statistics of two entities [2]. We compute PHI-COEFFICIENT as follows:

$$
\phi(q, r) = \frac{|Q \cap R| \sqrt{|Q||R|-|Q-(Q \cap R)||R-(Q \cap R)|}}{\sqrt{|Q||R||R||Q||Q||R||R|}},
$$

where $Q$ is the set of documents containing query entity $q$ and $R$ is the set of documents containing related entity $r$. We compute association scores $A(q, r)$ for every pair $(q, r)$ where $q$ and $r$ occur together in at least one document.

Computing associations based on document and sentence level co-occurrence will yield different results. Our working hypothesis is that co-occurrences at the document level results in more topical associations, whereas co-occurrences at the sentence level indicates a more relational or functional association. We consider both levels as important in our exploratory search task. Furthermore, co-occurrence at the document level could also help with co-reference problems, i.e., in a news article the first mention of a person is usually complete (with full name, title, etc), and then referenced only with the first or last name in subsequent mentions. To capture associations at both document and sentence level, we compute both association scores and combine them in a linear combination:

$$
A(q, r) = \lambda A_s(q, r) + (1 - \lambda) A_d(q, r)
$$

where $A_s$ denotes the association score at the sentence level, $A_d$ denotes the score at the document level, and $\lambda$ as the weighting parameter.

Temporal extent. Once we have computed associations between entities through co-occurrence, we proceed to identify the time period where these associations happen. We start with a naive approach: using the publication dates of the articles in which the first and last co-occurrences are found, respectively. A more refined method for defining this temporal extent is left to future work.

Association context. To give users a better idea of the context of associations between entities, we provide a summary of the documents containing the evidence. In contrast with relation extraction, where every pair of entities is labeled with a single relation label, we aim to provide several related terms. These terms represent a set of possibilities, in which topical associations and/or functional associations might be found. For our work-in-progress, we simply use the most frequent terms from the documents.
plain the associations visualized in the related entities panel. To help users understanding the associations between entities, the association context is visualized with a tag cloud. In every result, the title, date, a text fragment, and the full text of the document is displayed. The query and related entity are highlighted in the text fragment. The documents that are listed here can be ordered either by relevance score or by chronological order (document publication date). This chronological ordering will help users in going through the document collection once they discover an interesting document selection, thus allowing them view the document selection as a story.

Search interaction model. A user can gain insights from our interface in the following way. He starts an interaction by entering a query to the interface. This query will cause two things to change in the user interface. Firstly, the document frequency of the query entity over time will be displayed in the query entity histogram panel. Then, a set of related entities will be displayed in the related entities panel, alongside the timeline bar of the associations. Provided with these visual summaries, the user can quickly proceed to refine their query, either by selecting a time period, or by focusing on a relation. The user will get an idea about the context of the document selection through the evidence summary. The user can repeat this process until he has found a document selection of interest. Once the user has discovered this set, he can go through the documents in chronological order.

5. INITIAL EVALUATION

A full evaluation of our time-aware entity-association approach is part of ongoing work. In this section we report on case studies, involving three individual end users: a historian, a political scientist, and an anthropologist. In this initial evaluation, we are concerned with two aspects: (1) whether the interface actually help in search task, and (2) usability of user interface elements. We use a collection of 140,000 news articles spanning a 12 year period. These articles are news articles about Indonesia, the area of expertise of our subjects. The articles are published in national and international news media. One of the interesting features of this collection is that the articles are pre-selected manually by a group of editors and may have uneven temporal and topical distribution. Our users use this collection for various research purposes, including identifying the political landscape in Indonesia during a given time period, and extracting entity networks of political figures. Therefore, our exploratory search interface will also be used to get an overview of the quality and completeness of the collection.

In our evaluation, we distinguish between two main search tasks: profiling and discovery. We define profiling as the case where a user has sufficient background knowledge on a query entity, thus the goal of the search task is to identity whether the information inferred from the collection actually reflects this background knowledge. In discovery, we focus on whether the interface can point to an interesting document selection, and help discover new associations. In this task, users submit a query entity that they are less familiar with.

The profiling task. For the profiling task, our end users choose several queries that they are familiar with. One such query is “Prabowo Subianto,” a political figure in Indonesia. A first look at the query result shows an uneven temporal distribution of the documents mentioning “Prabowo Subianto.” His mentions increased particularly in 2009, but this rise actually started in 2008, when he is first considered as a presidential candidate (as shown in Figure 2).

A particularly interesting case of discovery is found in a related entity: “Husni Maderi,” who has strong associations with “Ani Yudhoyono” (shown in 3) in 2010. Upon discovering this association, our users turn to the evidence panel. From the evidence panel, especially the highlighted text snippets, we find out that Husni Maderi is the wife of Susilo Bambang Yudhoyono. In contrast with the previous query, our users do not have much background knowledge on this person, therefore allowing more chance discoveries. About half of the entities that are returned are interesting to the users. Most of them indicate close, but subtle ties based on family relationships with other political actors, a topic the users are particularly interested in.

Discovery task. In the discovery task, the users choose a query entity that they are less familiar with: “Ani Yudhoyono,” the wife of the current Indonesian president Susilo Bambang Yudhoyono. Our end users confirm that the previous scenario is where an exploratory interface might help them focus on an interesting document selection. By seeing related entities, the users are pointed to queries and document selections that they normally would not try in a general search scenario.

6. CONCLUSION AND DISCUSSION

We have presented an interface for exploring entity associations over time. In our preliminary evaluation, we let our users use the interface for profiling and discovery. We find that our end users can use the interface intuitively without much guidance. It provides an interactive way for users to engage with large document collections through entities and relationships. The entity-oriented approach helps the users by directing their attention to interesting document sets. This is a promising way to select documents for further analysis, such as network analysis, either qualitatively or quantitatively.

As to future work, we aim to provide a better explanation of the
association context. Identifying topical and relational terms around entity pairs would give a more useful summary in the evidence panel. Associations between two entities might evolve over time. In a historical or political domain that our target end users are interested in, this could mean two political figures connect through different channels. Investigating how this association context evolves over time, and visualizing it, is an interesting research direction.

Acknowledgements. This research was partially supported by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSIne project), the Netherlands Organisation for Scientific Research (NWO) under project nr 640.004.802, 727.011.005, 612.001.116, H08-11-10, the Center for Creation, Content and Technology (CCCT), the QaMerdes project funded by the CLARIN-nd program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105 and the Yahoo! Faculty Research and Engagement Program.

7. REFERENCES


