ploratory search for a particular time interval without providing tags as input. Another index allows retrieving all versions of a website that have been tagged during a given time period together with the tags (i.e., *UrlTagMapping*). For a compact index structure, two mappings assign ids to tags and websites, which are used exclusively in all other indexes and mappings (i.e., *IdTagMapping*, *IdUrlMapping*). Even though there is much room for improvement and optimization of temporal indexes [8, 9], the rather simple mappings, which can easily be constructed using a distributed data processing platform like *Hadoop*, already go a long way.

All indexes are ordered by their keys (i.e., ID, year, month or pair of tag and year or website and year, respectively). The fetching of the indexes is realized by web services, which are invoked separately for tags and websites as well as a single fetch for the versions with tags of each website. After fetching the data, all items (i.e., tags, websites or versions) are ordered according the ranking functions in Sec. 3.3.

The query with tags as well as the time period, consisting of start year and month as well as end year and month, can be entered and selected at the top of the page, using the search input and the three sliders as shown in Fig. 1.

4. CONCLUSIONS AND OUTLOOK

Tags have been proven to be a reasonable surrogate for the actual content of a website. As indexing complete Web archives in a full text index as well as providing temporal IR capabilities to this appears to be a big challenge, meta information, such as tags from social bookmarking services, are a good way to mimic the real content of different versions of a website. We have not yet evaluated the effectiveness of our method in terms of meeting a users information need, however, it is definitely a great improvement over accessing a Web archive by providing the exact URL and time of a website's version, like most Web archives only allow today.

The different levels of co-occurring tags to be considered as sub-topics of the original query in Tempas provide an easy way to explore the available dataset. The ability to specify a time period to search in enables temporal IR capabilities for Web archives. However, in order to present the retrieved websites to the user, in its current state, Tempas only passively requests the intended version of a website from the Internet Archive's Wayback Machine (cf. Sec. 3). Therefore, we do not know whether an archived version is actually available and we do not have access to data like the title of the website, which therefore cannot be included in the search results. In the context of the ALEXANDRIA project[7] we own a subset of this archive, comprising the complete archived part of the German Web, which we have full access to. In coming versions of Tempas we will integrate this to provide richer information to the user. It will also allow us to compute statistics of the retrieved result set, such as the fraction of the actually archived Web that is covered by Tempas as well as the time gaps between tagged and archived versions of the websites.

In future research we are going to look into more sophisticated retrieval as well as ranking models. Also different visualizations and ways to explore archives are subject for further investigation. While we have shown how tags can serve as surrogates for the content of a website in temporal IR, there are other IR related challenges that need to be tackled in the future in order to enable the same convenience for temporal Web archive search that we are used to in common search engines on the current Web. For instance, query suggestions and reformulations, which support the users in formulating their information needs, are not available for archives yet. This is mainly due to the temporal characteristic, which raises new challenges for theses tasks, as well as the lack of query logs, which are commonly used for this purpose. We will investigate how tags from social bookmarking services can serve as surrogates for query logs as well as other required data in temporal IR on Web archives as well. Furthermore, we are planning on incorporating different meta data than tags that is available in a temporal fashion or can be extracted from an archive as derivative dataset. This might include postings on social websites such as Twitter as well as host graphs of websites, which can be derived from the crawls of archived websites.

[7] `http://alexandria-project.eu`

References

[1] Susan Schreibman, Ray Siemens, and John Unsworth. *A Companion to Digital Humanities*. 2008.

[2] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, and Jon Orwant. Quantitative Analysis of Culture Using Millions of Digitized Books. *science*, 2010.

[3] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. Computational journalism: A call to arms to database researchers. In *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research*, pages 148–151, 2011.

[4] Ricardo Campos, Gaël Dias, Alípio Mário Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. 2014.

[5] Arkaitz Zubiaga, Victor Fresno, Raquel Martinez, and Alberto Perez Garcia-Plaza. Harnessing folksonomies to produce a social classification of resources. *IEEE Transactions on Knowledge and Data Engineering*, 25 (8):1801–1813, 2013.

[6] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 193–202, 2008.

[7] Raluca Paiu. *Exploiting Tag Information for Search and Personalization.* PhD thesis, Leibniz Universität Hannover, 2009.

[8] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index maintenance for time-travel text search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, .

[9] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Efficient temporal keyword search over versioned text. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, .