

May Require Greece to Cut Budget and Jobs’, ‘I.M.F. Is Playing the Role of Deal Maker in Europe’, ‘E.U. Leaders Turn to I.M.F. Amid Financial Crisis’. After *decreasing* the importance of ‘Greece’, a new line appeared, focusing on the *Spanish* economic struggle. Representative articles include ‘Spain Seen as Moving Too Slowly on Financial Reforms’ and ‘I.M.F. Gives Backing to Spain’s Austerity Measures’.

7. RELATED WORK

To the best of our knowledge, the problem of constructing metro maps automatically is novel. There has been extensive work done on related topics from topic detection and tracking to summarization and temporal text mining.

Our work differs from previous work in two important aspects. Our system has **structured output**: Not only does our system pick nuggets of information, it explicitly shows connections among them. Prior work, in contrast, has been limited largely to list-output models. In the summarization task [14, 2, 15], the goal is often to summarize a corpus of texts by extracting a list of sentences. Other methods [10, 20, 19] discover new events, but do not attempt to string them together.

Numerous prior efforts have moved beyond list-output, and proposed different notions of storylines [1, 17, 18, 2]. Graph representations are common across a variety of related problems [9, 7, 12], from topic evolution to news analysis. However, in all of those methods, there is no notion of **path-coherence**. In other words, the edges in the graph are selected because they pass some threshold, or belong to a spanning tree. We believe that the notion of coherent paths facilitates the process of knowledge acquisition for the users.

Finally, different notions of coherence and coverage have been proposed in the literature. For example, enhancing coverage has been explored in the context of ranking and summarization (see MMR [21]). We chose not to use MMR as it does not provide approximation guarantees, and could not be combined with our orienteering algorithm. Modeling *coherence* via lexical relations was studied in [3]. However, their notion is restricted to chains of related words (Machine, Microprocessor, Device). In contrast, we generate coherent chains of *articles* by taking multiple concepts into account.

8. CONCLUSIONS AND FUTURE WORK

We have presented a new task, creating structured summaries of information, which we call *metro maps*. Given a query, our algorithm generates a concise structured set of documents which maximizes coverage of salient pieces of information. Most importantly, metro maps explicitly show the relations between the retrieved pieces.

We formalized the characteristics of good metro maps and provided efficient methods with theoretical guarantees. Our approach finds concise maps, making it well-suited for complement existing visualization and user interaction approaches. In particular, we integrate user preferences into our framework by providing an appropriate user-interaction model based on feature-based feedback.

We conducted pilot user studies, testing our algorithm on a real-world dataset. The study showed the promised of the proposed approach as an effective and fast method for creating valuable metro maps.

In the future, we plan to pursue richer forms of input, output and interaction, and the incorporation of higher-level semantic relations into our framework. In addition, we would like to apply our methods to other datasets, such as scien-

tific publications. We believe that metro maps will enable users to better cope with information overload.

Acknowledgments: This work was partially supported by ONR PECASE N000141010672, ARO MURI W911NF0810242, and NSF Career IIS-0644225. Dafna Shahaf was supported in part by Microsoft Research Graduate Fellowship.

9. REFERENCES

- [1] A. Ahmed, Q. Ho, J. Eisenstein, E. Xing, A. J. Smola, and C. H. Teo. Unified analysis of streaming news. In *WWW'11*, 2011.
- [2] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR '01*, 2001.
- [3] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *ACL Workshop on Intelligent Scalable Text Summarization*, 1997.
- [4] C. Chekuri and M. Pal. A recursive greedy algorithm for walks in directed graphs. In *FOCS '05*, 2005.
- [5] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *SIGIR '08*, pages 595–602. ACM, 2008.
- [6] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD '09*, 2009.
- [7] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *KDD '04*, 2004.
- [8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, June 1981.
- [9] Y. Jo, J. E. Hopcroft, and C. Lagoze. The web of topics: discovering the topology of topic evolution in a corpus. In *WWW '11*, 2011.
- [10] J. Kleinberg. Bursty and hierarchical structure in streams, 2002.
- [11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.
- [12] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM '04*, 2004.
- [13] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14, 1978.
- [14] A. Nenkova and K. McKeown. A survey of text summarization techniques. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*. 2012.
- [15] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. Newsinessence: summarizing online news topics. *Commun. ACM*, 48:95–98, October 2005.
- [16] D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *KDD '10*, pages 623–632, New York, NY, USA, 2010. ACM.
- [17] R. Swan and D. Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *KDD '00*, 2000.
- [18] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *SIGIR '11*, 2011.
- [19] Y. Yang, T. Ault, T. Pierce, and C. Lattimer. Improving text categorization methods for event tracking. In *SIGIR '00*, 2000.
- [20] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4), 1999.
- [21] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*. ACM, 2003.