Summarization of Changes in Dynamic Text Collections

Manika Kar

PhD in Computer Science (MAP-i), INESC TEC, Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
manika.kar@fe.up.pt

Information Retrieval is the Informatics field primarily focused on all problems and challenges related to information storage and access. The large majority of works in this area are based on static collections of documents. However, many of these collections are dynamic, and have evolved over time with documents being added, edited or simply removed at different times. Even in highly dynamic environments such as the World Wide Web, research tends to be centered on the most recent version of the documents and all the past information is normally discarded. Recognizing these changes over dynamic text collections and exploiting them for document retrieval and presentation purposes introduce new and relevant research challenges. This paper addresses the opportunity that gains relevance in this context - summarization of changes in dynamic text collections. We first define the problem in order to produce a summary that describes textual changes to an entire document or a set of related documents over an user defined time period. Then, from literature we present an extensive overview of the relevant approaches depicting similar problems and at last some discussions including future aspects.

Changes summarization, Update summarization, Dynamic text collections

1. INTRODUCTION

World Wide Web or web (in short) has become the biggest information repository in the world. Web page contents change rapidly over time due to the low cost of publishing information on the web. Several studies have focused and documented this dynamic nature of the web (Fetterly et al. 2004; Adar et al. 2009). The changes of web page contents can be of different sizes. Usually, either some new information is added or old text is deleted in addition to some unchanged, static content in the page. The ultimate case occurs when the whole page is deleted or a new one is created. Sometimes, these changes are related to reflect real-world events but in other cases the changes are only generic i.e. just to maintain the document contents properly. Recognizing these changes and exploiting them for document retrieval and presentation purposes introduce new research challenges. In this context, the summarization of changes in dynamic text collections is a relevant task.

In traditional summarization, retrieval techniques are used to produce a summary to catch the major points expressed in an entire document (or collection of documents) and present them in a condensed form to the user. The problem of traditional summarization

has become much popular since the idea started by Luhn (1958). This problem has been studied in many variations through a vast number of summarization techniques (Erkan and Radev 2004b; Harabagiu and Lacatusu 2005; Shen et al. 2007; Mani 2001). But most of their works have focused on static document collections, without attempting to find the most important changes from different variations of information on a specific topic.

Unlike traditional summarization, the idea behind summarization of changes is to produce a concise, informative summary that describes textual alterations to a document or a set of related documents between a certain time interval. In other words, the aim is on the development of tools that are able to produce an automatic summary of an user-defined period in the lifetime of a dynamic text collection. Users can be overwhelmed more often with available enormous information in their area of curiosity over a certain time period. It may be difficult to track which of them contain major changes. Thus, instead of producing a generic summary, a summary describing the major changes to a document or a collection of documents can be of some value in several situations. For example, a journalist wants to know an overview of main events, (on a specific topic) for which all the modifications were made

to a set of documents last week. Thus saving the user from browsing the web content during a long period. Overall, summarization of changes would allow users to answer questions such as:

- What were the important changes that have occurred in a collection of documents over a time period?
- What is the summary of the changes made to a document between two different revisions?
- Which were the important events that have occurred during an user-defined period?

Not only news articles but also Wikipedia, the collaboratively edited encyclopedia available on the web, is a major example of dynamic text collections. Wikipedia is a pertinent resource for research in the summarization of changes task due to two main reasons. First, the entire revision history of every web page is kept and these revisions can be accessed publicly through an API. Second, others are therefore able to reproduce easily on someone's findings.

On the other hand, online social networks, such as Facebook or Twitter or blogs, are another examples of time dependent collections. Summarization of changes is also useful in these media as they are distinguished by the temporal dynamics of the content as well as high posting volume.

2. RELATED WORK

The idea, summarization of changes is mainly motivated by WikiChanges (Nunes et al. 2008). The authors observe that the distribution of the revision history of any Wikipedia article over time is highly correlated with the article's popularity. Following this observation the task, summarization of changes is identified. Our idea builds on several areas of research.

Document Understanding Conference (DUC)¹ launched update summarization² as a pilot task in 2007. This task has become much popular in the following years, as can be noticed by the growing number of participants to this special track organized by DUC and Text Analysis Conference (TAC). The task is focused on the generation of an update summary for multiple documents at current time based on a common topic under the assumption that the user knows a set of past documents. The purpose of each update summary will be to inform the reader with novel information on that specific topic. We can say that this is closely similar with our idea pointing out only the changes while the

¹<http://duc.nist.gov/>

²<http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html>

common parts are already known to the reader. But, one of the main differences between update summarization and summarization of changes is the information need being addressed is restricted to "current" changes in an update summary. Where as we consider that summarization of changes is not limited to current changes and should address any given time period.

Within the topic detection and tracking (TDT) framework, in particular summarization of changes is related to first story detection and tracking tasks depicted by Allan (2002). However, the aim of first story detection task is only to classify some stories as a new (or old), but not to describe this with respect to time. Summarization of changes builds on another area of research, the "bursty" event detection (Neil and Wong 2009) as most of the times the major changes happen due to the occurrences of "bursty" events. Our idea is related to other tasks such as novelty detection (Soboroff and Harman 2005), which can be defined as the problem of determining whether a document contains new information given an existing collection. Thus, while the goal of novelty detection is to determine whether some information is new, the goal of summarizing changes is to extract and synthesize the changed information. Summarization of changes is also related to contrastive summarization, i.e. the problem of jointly generating summaries for two entities in order to emphasize their differences (Lerman and McDonald 2009). Throughout this paper, our discussion focuses on different extractive summarization techniques mostly relevant in the context of our task.

2.1. Term Frequency Based Methods

The term-weighting measure based on raw term frequency is a core technique in extractive summarization. The explicit reference towards the idea summarization of changes is Jatowt et al. (2004). The authors proposed ChangeSummarizer, which can produce summaries of textual changes in web collections. ChangeSummarizer helps users in searching for new relevant information in their interest by providing the summary of recent, important changes related to that topic. Each web page is compared between the old and new web collections. After comparison, the new terms are extracted. The system then calculates scores for each term according to the popularity of the term in static and dynamic parts of the collection based on term frequencies. Finally the system presents the sentences with highest overall scores.

Nunes et al. (2008) present WikiChanges, a web-based application designed to plot Wikipedia article's revision history in real time and to produce a temporal summary. The summary addresses what changes occurred during a given set of revisions. A

very simple approach is used based on the terms inserted between a start and end revisions (all intermediary revisions are ignored) of an article. Each term is then scored by subtracting the old terms frequency count from the new terms frequency count. The final top scored terms can be presented to the final user or used as input to a sentence selection algorithm. This is the first approach to the revisions summarization task using tag clouds to present an automatic summary.

2.2. Topic Model Based Methods

As discussed before, the changes summarization task is closely related to update summarization but in a more general way. That's why we intend to explore the methods used for update summarization or other relevant tasks.

Topic Model based methods used explicitly in update summarization task are basically variations of Latent Dirichlet Allocation (LDA) (Blei et al. 2003) and aim to learn to distinguish between common and new information. Previous Bayesian works for traditional summarization (Daumé III and Marcu 2006; Chemudugunta et al. 2006; Haghighi and Vanderwende 2009) only learn collection-specific distributions to capture the main topic in the collection. To identify and extract novel information comparing between base (the collection of earlier documents) and update (the collection of recent documents) pairs, a nonparametric Bayesian approach DUALSUM (Delort and Alfonseca 2012) is specifically designed. The main idea in DUALSUM comes from TOPICSUM (Haghighi and Vanderwende 2009), with two main changes. First, TOPICSUM can only learn the main theme of a collection, DUALSUM focuses on the differences between two collections. Second, TOPICSUM has an additional layer to model topic distributions at the sentence level, in DUALSUM this assumption is removed and the topic distribution is modeled at document level.

In DUALSUM, documents are modeled as a bag of words that are assumed to be sampled from a mixture of latent topics. For a document in the base collection, words can be originated from one of three different topic distributions: a general background distribution over common words, a document-specific distribution for each document in the collection pair (base and update) and the common information between the both collections, in which the last one captures the main theme. For a document in the update collection, words can be associated with one of the four different topic distributions: a general background distribution over common words, a document-specific distribution for each document in the collection pair (base and update), the common information between both collections and the topic distribution of update

collection, the last one captures the most important changes to the main theme.

2.3. Graph Based Methods

Graph-based ranking algorithms, such as Google's PageRank (Brin and Page 1998) and Kleinberg's HITS algorithm (Kleinberg 1999) have been traditionally and successfully used in the analysis of link-structure of the World Wide Web, citation analysis and social networks. A similar graph-based ranking algorithm called LexRank is developed for calculating sentence significance (Erkan and Radev 2004a,b). In their algorithm the documents are represented as a weighted undirected graph considering sentences as vertices and cosine similarity between sentences as the edge weight function. Mihalcea (2004) describes another graph based ranking algorithm for automatic sentence extraction in the context of text summarization.

In recent years, graph based ranking techniques are also deployed for update summarization and are becoming more and more popular nowadays. PNR² (Ranking Sentences with Positive and Negative Reinforcement) (Wenjie et al. 2009) and MRSP (Manifold ranking with sink points) (Du et al. 2009) are two such popular methods. These methods construct a unified sentence graph from base and update collections, in which reinforcements between sentences are used to decide their scores and the highest scored sentences are pulled out to form a summary. PNR² uses negative reinforcements to PageRank, where as MRSP uses reinforcement losses for Manifold Ranking. The base collection take part directly in the reinforcement propagation of the update collection. As reinforcement propagation is applied to decide sentence salience (see Erkan and Radev 2004b), one following problem is that salience of the sentences in update collection may be disturbed by base collection. To overcome this problem Li et al. (2013) introduce another graph-ranking based method. The method uses constrained reinforcements on a sentence graph, which unifies base and update collection to determine the salience of the sentences. In general, the complexity of this kind of approaches is very high. To overcome the problem, the authors proposed an approximate method (QPSum) that can solve the problem in polynomial time. For ranking the graph, PNR² is used as an extension of the TextRank algorithm (Mihalcea and Tarau 2004) for update summarization. The ranking method constructs the following matrix

$$\mathbf{M} = \begin{pmatrix} \alpha_1 \mathbf{W}_{AA} & \beta_1 \mathbf{W}_{AB} \\ \beta_2 \mathbf{W}_{BA} & \alpha_2 \mathbf{W}_{BB} \end{pmatrix}$$

, where $\alpha_1, \alpha_2 > 0$ and $\beta_1, \beta_2 < 0$ and $\mathbf{W}_{AB}$ indicates the similarity matrices between sentences

A and B . Here, $\alpha_1 \mathbf{W}_{AA}$ and $\alpha_2 \mathbf{W}_{BB}$ are to execute positive reinforcements and similarly, $\beta_1 \mathbf{W}_{AB}$ and $\beta_2 \mathbf{W}_{BA}$ are to execute negative reinforcements. After normalizing $\mathbf{M}$ by column, finally ranking scores $\mathbf{f}$ are calculated with the equation $(\mathbf{I} - \theta \cdot \mathbf{M}) \cdot \mathbf{f} = \mathbf{p}$, where $0 \leq \theta \leq 1$.

The method is applied on the TAC 2008 and 2009 benchmark data sets which consist of a number of topics, each associated with two document collections say, A and B . For each document and each topic, OpenNLP tool³ is applied to detect and tokenize sentences. After removing stop words in both documents, the sentences are represented as vectors of words containing a boolean value. Cosine distance is used to measure topic relevance and sentence similarity. The performance of the method is evaluated through ROUGE⁴, which is officially adopted by TAC for evaluation of automatic generated summaries.

2.4. Clustering Based Methods

Clustering based summarization methods are also another popular techniques for generic summarization. These methods usually apply different clustering techniques on the term-sentence matrices formed from the documents. After the sentences are grouped into different clusters, a centroid score is assigned to each sentence based on the average cosine similarity between the sentence and the rest of the sentences in the same cluster. Finally, the sentences with the highest score from each cluster are selected to form the summary.

Wang and Li (2010) proposed an incremental hierarchical clustering based approach to update document summaries in real time when new documents arrive. The COBWEB algorithm originally was built by Fisher (1987), is applied here to build a sentence hierarchical tree. When a new element comes, the COBWEB algorithm traverses the tree top-down fashion starting from the root. During traversing, the COBWEB algorithm executes one of the four possible operations (*insert*, *create*, *merge* and *split*) based on maximizing the criterion function.

Recently Georgescu et al. (2013) have presented Wikipedia Event Reporter, a system that automatically extracts events from the Wikipedia revision history and presents related information summary to the user. A change in Wikipedia is meant by the updates occurred in one revision when compared to the previous revision of an article. Each revision of an article has the changes along with its creation time, its author, and, possibly, comments given by the updater. The first step in Wikipedia Event Reporter is to identify bursts for any article using the burst detection algorithm by Zhu and Shasha (2003). The

algorithm applies a sliding time window for which the number of changes is counted. The corresponding time intervals for which the change rate exceeds a certain threshold are considered as burst. In the next step, a classifier classifies between “event-related” and “not event-related” burst. Finally, event-related information are summarized through clustering of changes depending on various types of information such as change time, textual similarity, and the position of edits within an article.

2.5. Simple Filtering Rules

To capture the changed information from current documents comparing to the earlier documents, the first challenge is to filter the redundant information. Zhang et al. (2009) proposed three filtering strategies: document filtering, summary filtering and union filtering. These strategies are based on the degree of membership from fuzzy set theory to measure the similarity of sentences between earlier and current information. After that, the filtered sentences are ranked through two approaches. The first one is a signature based approach, in which temporal topic signatures are extracted from the filtered sentences and the second one is manifold ranking based approach, in which the macro-structure of the filtered sentences can be reserved. Schilder et al. (2008) apply regression Support Vector Machine as a filter, which can extract sentences that resemble first sentences in the entire news articles. The intuition behind this idea is like first sentences have less anaphoric expressions. After extracting the sentences a modified version of FastSum (Schilder and Kondadadi 2008) is applied.

2.6. Other Methods

Other methods include first temporal summarization approached by Allan et al. (2001) to monitor the changes of news streams over time. This type of work aims to catch new information during the evolutionary change of news articles. Another work is the STORIES system (Subašić and Berendt 2010) on news streams, which is very similar towards the intention of our idea. The authors presented the STORIES system for learning an abstracted story from dynamic news collection, making a story graph for story tracking and describing the summary from a selected time period. The method for story understanding is built on multiple co-occurrences within one document. While the work, Discriminative Sentence Selection (DSS) (Wang et al. 2012) based on a multivariate normal model aims to summarize the differences among different document groups. Other works related with Wikipedia revisions history (Keegan et al. 2012; Whiting et al. 2012; Steiner et al. 2013) have focused on mainly detecting events instead of generating temporal summary

³<http://opennlp.apache.org/>

⁴<http://www.isi.edu/publications/licensed-sw/BE/index.html>

automatically for event-related information.

Within time-dependent collections there has been prior works on understanding the temporal evidence of the entire blogosphere (Chi et al. 2006; Mei et al. 2006). To our best knowledge, the only explicit reference is Lin and Sundaram (2007), introducing temporal summarization over four different blog types. The authors proposed a framework to expose how the themes evolve over time based on non-negative self-similarity matrix factorization and the blog antenna visual metaphor is to summarize the blog temporal dynamics.

3. DISCUSSION

Summarization of changes unlike most summarization research is focused on summarizing changes in dynamic text collections. Previous work on this particular problem is relatively scarce. Most of the related works are centered on summarizing “current, important changes”, but do not define summarizing changes broadly i.e., within any user defined period. So, there is a need to develop new methods and techniques in future to summarize changes in a more general time frame. An important challenge for this task is to identify event-related changes and discard generic modifications in highly active dynamic collections. The relevant approaches to similar problems discussed in the previous section only considers the newest and the oldest revision of an article, ignoring the intermediate revisions. The effectiveness of the summarization incorporating those integral revisions is another unexplored challenge. In spite of these challenges, the current existing issues in traditional summarization are present here to solve. Overall, as another step towards summarization, summarization of changes research needs further study.

ACKNOWLEDGEMENTS

I would like to sincerely thank my PhD supervisor Sérgio Nunes and co-supervisor Cristina Ribeiro for giving their valuable comments to significantly improve each version of this article. I would also like to thank three anonymous reviewers, whose feedback was very useful to make the final version. This work is financially supported by Erasmus Mundus Mover project.

REFERENCES

Fetterly, D., Manasse, M., Najork, M. and Wiener, J. L. (2004) *A large-scale study of the evolution of web pages*. Softw. Pract. Exper., 34(2), 213-237.

Adar, E., Teevan, J., Dumais, S. T. and Elsas, J. L. (2009) *The web changes everything:*

understanding the dynamics of web content. Proc. of the 2nd Int. Conf. on WSDM'09, pp. 282-291.

- Luhn, H. P. (1958) *The automatic creation of literature abstracts*. IBM J. of Res. and Dev., 2(2), 159-165.
- Erkan, G. and Radev, D. R. (2004) *LexRank: Graph-based Centrality as Salience in Text Summarization*. J. of Art. Int. Res. (JAIR), 22(1), 457-479.
- Harabagiu, S. and Lacatusu, V. F. (2005) *Topic themes for multi-document summarization*. Proc. of SIGIR'05, pp. 202-209. ACM.
- Shen, D., Sun, J. T., Li, H., Yang, Q. and Chen, Z. (2007) *Document summarization using conditional random fields*. Proc. of the 20th Int. joint conf. on Artificial intelligence'07, pp. 2862-2867.
- Mani, I. (2001) *Recent developments in text summarization*. Proc. of CIKM'01, pp. 529-531. ACM.
- Nunes, S., Ribeiro, C. and David, G. (2008) *Wikichanges - exposing Wikipedia revision activity*. Proc. of the 4th Int. Symp. on Wikis, WikiSym'08.
- Allan, J. F. (2002) *Topic detection and tracking*. Springer, Berlin.
- Neil, D. and Wong, W. (2009) *Tutorial on event detection KDD 2009*.
- Soboroff, I. and Harman, D. (2005) *Novelty detection: the trec experience*. Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing, HLT'05, pp. 105-112. ACL.
- Lerman, K. and McDonald, R. (2009) *Contrastive summarization: an experiment with consumer reviews*. Proc. of NAACL-Short09, pp. 113-116. Association for Computational Linguistics (ACL).
- Jatowt, A., Bun, K. K. and Ishizuka, M. (2004) *Change summarization in web collections*. Proc. of the 17th Int. Conf. on Innovations in applied artificial intelligence, pp. 653-662. Springer.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) *Latent dirichlet allocation*. J. Mach. Learn. Res.(JMLR), 3, 993-1022.
- Daumé III, H. and Marcu, D. (2006) *Bayesian Query-Focused Summarization*. Proc. of the 21st Int. Conf. on Computational Linguistics and the 44th annual meeting of the ACL, pp. 305-312, ACL.
- Chemudugunta, C., Smyth, P. and Steyvers, M. (2006) *Modeling general and specific aspects of documents with a probabilistic topic model*. In NIPS, pp. 241-248.

- Haghighi, A. and Vanderwende, L. (2009) *Exploring content models for multi-document summarization*. Proc. of Human Language Technologies: The Annual Conf. of the North American Chapter of the ACL, pp. 362-370. ACL.
- Delort, J. Y. and Alfonseca, E. (2012) *DualSum: a Topic-Model based approach for update summarization*. Proc. of the 13th Conf. of the European Chapter of the ACL, pp. 214-223. ACL.
- Brin, S. and Page, L. (1998) *The Anatomy of a Large-scale Hypertextual Web Search Engine*. Computer Networks and ISDN Systems, 30(1-7), 107-117.
- Kleinberg, J. M. (1999) *Authoritative Sources in Hyperlinked Environment*. Journal of the ACM, 46(5), 604-632.
- Erkan, G. and Radev, D. R. (2004) *LexPageRank: Prestige in Multi-Document Text Summarization*. Proc. of EMNLP, pp. 365-371.
- Mihalcea, R. (2004) *Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization*. Proc. of the ACL 2004 on Interactive poster and demonstration sessions, ACL.
- Wenjie, L., Furu, W., Qin, L. and Yanxiang, H. (2009) *ranking sentences with positive and negative reinforcement for query-oriented update summarization*. Proc. of the 22nd Int. Conf. on Computational Linguistics, pp. 489-496. ACL.
- Du, P., Guo, J., Zhang, J. and Cheng, X. (2009) *Manifold ranking with sink points for update summarization*. Proc. of CIKM'10, pp. 1757-1760.
- Li, X. and Du, L. and Shen, Y. D. (2013) *Update Summarization via Graph-Based Sentence Ranking*. IEEE Trans. On Knowledge and Data Engineering, 25(5), 1162-1174.
- Mihalcea, R. and Tarau, P. (2004) *TextRank: Bringing Order into Texts*. Proc. of EMNLP'04, pp. 404-411. ACL.
- Wang, D. and Li, T. (2010) *Document Update Summarization Using Incremental Hierarchical Clustering*. Proc. of the 19th ACM Int. Conf. on Information and knowledge management, pp. 279-288. ACM.
- Fisher, D. H. *Knowledge acquisition via incremental conceptual clustering*. Machine Learning, 2, 139-172.
- Georgescu, M., Pham, D. D., Kanhabua, N., Zerr, S., Siersdorfer, S. and Nejd, W. (2013) *Temporal Summarization of Event-Related Updates in Wikipedia*. Proc. of the 22nd Int. Conf. on WWW companion, pp. 281-284.
- Zhu, Y. and Shasha, D. (2003) *Efficient elastic burst detection in data streams*. Proc. of KDD'03, pp. 336-345. ACM.
- Zhang, J., Du, P., Xu, H. and Cheng, X. (2009) *ICTGrasper at TAC2009: Temporal Preferred Update Summarization*. Proc. of TAC'09.
- Schilder, F., Kondadadi, R., Leidner, J. L. and Conrad, J. G. (2008) *Thomson reuters at tac 2008: Aggressive filtering with fastsum for update and opinion summarization*. Proc. of TAC'08.
- Schilder, F. and Kondadadi, R. (2008) *FastSum: Fast and Accurate Query-based Multi-document Summarization*. Proc. of ACL'08: HLT, Short Papers, pp. 205-208. ACL.
- Allan, J., Gupta, R. and Khandelwal, V. (2001) *Temporal summaries of new topics*. Proc. of the 24th SIGIR'01, pp. 10-18. ACM.
- Subašić, I., and Berendt, B. (2010) *Experience STORIES: A Visual News Search and Summarization System*. Proc. of ECML PKDD'10, pp. 619-623, Springer-Verlag.
- Wang, D., Zhu, S., Li, T. and Gong, Y. (2012) *Comparative Document Summarization via Discriminative Sentence Selection*. ACM Trans. on Knowledge Discovery from Data, 6(3).
- Keegan, B., Gergle, D. and Contractor, N. (2012) *Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations*. Proc. of WikiSym'12, pp. 1. ACM.
- Whiting, S., Zhou, K., Jose, J. M., Alonso, O. and Leelanupab, T. (2012) *Crowdtiles: presenting crowd-based information for event-driven information needs*. Proc. of CIKM'12, pp. 2698-2700.
- Steiner, T., Hooland, S. V. and Summers, E. (2013) *MJ no more: Using Concurrent Wikipedia Edit Spikes with Social Network Plausibility Checks for Breaking News Detection*. Proc. of the 22nd Int. Conf. on WWW companion'13.
- Chi, Y., Tseng, B. L. and Tatemura, J. I. (2006) *Eigentrend: trend analysis in the blogosphere based on singular value decompositions*. Proc. of CIKM'06, pp. 68-77. ACM.
- Mei, Q., Liu, C., Su, H. and Zhai, C. (2006) *A probabilistic approach to spatiotemporal theme pattern mining on weblogs*. Proc. of the 15th Int. Conf. on WWW, pp. 533-542. ACM.
- Lin, Y. and Sundaram, H. (2007) *Blog Antenna: Summarization of Personal Blog Temporal Dynamics Based on Self-Similarity Factorization*. Proc. of Multimedia and Expo'07, pp. 540-543. IEEE.