Information Processing and Management 51 (2015) 809–833

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Summarization of changes in dynamic text collections using Latent Dirichlet Allocation model

Manika Kar*, Sérgio Nunes, Cristina Ribeiro

INESC TEC, DEI, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n. 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 22 September 2014
Revised 26 May 2015
Accepted 1 June 2015
Available online 30 June 2015

Keywords:
Changes summarization
Temporal term weighting
Sentence ranking
Latent Dirichlet Allocation
Wikipedia

ABSTRACT

In the area of Information Retrieval, the task of automatic text summarization usually assumes a static underlying collection of documents, disregarding the temporal dimension of each document. However, in real world settings, collections and individual documents rarely stay unchanged over time. The World Wide Web is a prime example of a collection where information changes both frequently and significantly over time, with documents being added, modified or just deleted at different times. In this context, previous work addressing the summarization of web documents has simply discarded the dynamic nature of the web, considering only the latest published version of each individual document. This paper proposes and addresses a new challenge - the automatic summarization of changes in dynamic text collections. In standard text summarization, retrieval techniques present a summary to the user by capturing the major points expressed in the most recent version of an entire document in a condensed form. In this new task, the goal is to obtain a summary that describes the most significant changes made to a document during a given period. In other words, the idea is to have a summary of the revisions made to a document over a specific period of time. This paper proposes different approaches to generate summaries using extractive summarization techniques. First, individual terms are scored and then this information is used to rank and select sentences to produce the final summary. A system based on Latent Dirichlet Allocation model (LDA) is used to find the hidden topic structures of changes. The purpose of using the LDA model is to identify separate topics where the changed terms from each topic are likely to carry at least one significant change. The different approaches are then compared with the previous work in this area. A collection of articles from Wikipedia, including their revision history, is used to evaluate the proposed system. For each article, a temporal interval and a reference summary from the article's content are selected manually. The articles and intervals in which a significant event occurred are carefully selected. The summaries produced by each of the approaches are evaluated comparatively to the manual summaries using ROUGE metrics. It is observed that the approach using the LDA model outperforms all the other approaches. Statistical tests reveal that the differences in ROUGE scores for the LDA-based approach is statistically significant at 99% over baseline.

© 2015 Elsevier Ltd. All rights reserved.

* Corresponding author.
E-mail addresses: manika.kar@fe.up.pt (M. Kar), ssn@fe.up.pt (S. Nunes), mcr@fe.up.pt (C. Ribeiro).

http://dx.doi.org/10.1016/j.ipm.2015.06.002
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In the area of Information Retrieval, it is recognized that retrieval from dynamic text collections on the web brings several new research challenges (Allan, Croft, Moffat, & Sanderson, 2012). Web pages are continually added, removed, or edited, resulting in active collections of documents that are always being modified. It is common to observe a high rate of changes as a consequence of the occurrence of real-world events. However, there are also modifications to documents which are generic, namely those resulting from minor revisions or additions/modifications of outdated information. This paper addresses a problem that gains relevance in this context—the summarization of changes.

The summarization of changes can be described as follows: given an evolving document collection and a temporal period, generate a summary of significant alterations made to the collection of documents during that period. Wikipedia is a prime example of a dynamic collection, with clearly identified documents—the articles, whose evolution in time can be seen in the revision history. When searching "Pete Seeger" on Wikipedia, for example, it is possible to access the history of collaborative editing for the corresponding article, as all revisions of this article are stored. In order to obtain the summary of changes to "Pete Seeger" within the time range "January 2014", we would expect to obtain "Pete died in New York City on January 27, 2014". Even though it is true that "he was an American folk singer and activist", this summary expresses a more general and more static information which, in a context of summarization of changes, does not pertain to the period "January, 2014".

This paper considers the task of generating a summary of the changes in a set of documents. The set of documents may include different documents on the same topic originated in a collection, or a series of versions of the same document. The following three properties are expected from a summary of the changes:

- Time-dependency. The summary is expected to highlight the information that has been changed on the set of documents between two points in time. However, it should also exclude the static information existing in the documents.
- Significance. During a given time period, changes to the text take place for different reasons. Often, changes are not very significant, such as the correction of syntax or grammatical errors, the modification of links or changes regarding a past time period. This outdated information seems to be updated simultaneously when an event draws attention to a Wikipedia article. However, these updates do not focus on the reason for those changes within that particular time period. Hence, the challenge is to identify the meaningful information which has the potential to be a significant part of the summary, besides all other unnecessary changed texts for the given period. Irrelevant details do not belong in the summary.
- Non-redundancy. The summary is expected to be synthetic and therefore avoid redundant information. Two similar sentences carrying the same information should not be selected simultaneously.

The summarization of changes is a task that can have any kind of time-dependent text collections as input. Research has been conducted in the past on related tasks. The concept of *monitoring changes* (Allan, Gupta, & Khandelwal, 2001) has been proposed in 2001. This concept focuses on unstructured dynamic text collections such as news articles. The goal of *monitoring changes* is to keep users informed by extracting the main events from a stream of news stories on the same topic. Similarly, the task of summarizing changes can be used in news articles to generate a summary of changes within a given time period. Given a collection of document groups, *Comparative Extractive Document Summarization* (CDS) Wang et al. was proposed to generate a summary of the differences among these document groups sharing a similar topic. If these document groups have evolved over time, the goal is to generalize the CDS problem, extending it to the evolution of the differences over time. The result will then be a summary of the differences among time-dependent comparable document groups for a selected time period. Moreover, the summarization of changes can easily be adapted to other frameworks.

Along with news articles and document groups, Wikipedia is another example of a collection of documents where the data is dynamic by nature. Wikipedia is being continuously updated and maintained by a community of editors. Often, new contents originate from recent events such as sports competitions, political controversies, new research outcomes, awards, resignations, births, deaths or natural calamities. Modifications are also made when existing contents are revised. Other large dynamic collections are generated by social media, such as Facebook, Twitter or blogs. Social networks are very popular and people use them to provide status updates, share opinions or broadcast news. These media are characterized by a strong temporal dynamics of the content and a high posting volume.

The summarization of changes can be applied in several scenarios, one of them being search. When queries include temporal information, summarization can provide more focused snippets for search results. Query-oriented update summarization is another use for the summarization task in dynamic text collections. It poses new challenges to the sentence-ranking algorithms as it requires not only important and query-relevant information to be identified, but also novelty (Wenjie, Furu, Qin, & Yanxiang, 2009) to be captured. ChangeDetect has been proposed to detect a list of changes made to a user-given web page, and to notify the users via email. This allows web users to track the important changes made to their favorite web pages. Because it is impossible to see every possible change, with this service the users can pay attention to the details only when a summary of significant changes triggers enough interest. A similar situation occurs in enterprise and public environments, where information is always being updated into the existing shared repositories. This summary will make people aware of the changes in a concise form, either on a daily or weekly basis. A summary of changes can also be very useful for a journalist or a student exploring historical information that is no longer available in the current version of the documents on a specific topic. The summarization of changes can play an important role in online social networks. On Twitter

Download English Version:

https://daneshyari.com/en/article/514959

Download Persian Version:

https://daneshyari.com/article/514959

Daneshyari.com