A Binary Tree based Approach for Time based Page Ranking in Search Engines

Sabiha Firdaus[1], Kashfia Sailunaz[2] and Ahmed Shoeb Al Hasan[3]

[1] Bangladesh University of Business and Technology

Received: 12 December 2014 Accepted: 5 January 2015 Published: 15 January 2015

Abstract

Search engines rank web pages according to different conditions. Some of them use publication time, some use last time of update, some checks the currency of the content of the web page. In this paper, a new algorithm is proposed which will work on the time of the web page, temporal information of the content and forms a binary tree to rank among web pages.

Index terms—

1 Introduction

emporal information of web pages is normally collected from the Meta data or publication date of the web page. Sometimes they are extracted from the contents of the web page. Our idea is to find the temporal information from a webpage (from both publication date and its content) and use them to create a page ranking approach for search engines. The ranking of the web pages will be based on the temporal information of the web pages related to the query. The web pages containing the oldest information about the query topic should be on the top k-results.

2 II.

3 State of the Arts

As per Alonso [1], Temporal information are well-defined. They can be normalized and organized hierarchically. The temporal information can be any Date (e.g. January), Time (e.g. 3 p.m.), Duration (e.g. 3 years) or Set (e.g. twice a week). Temporal expressions of a query or web page can be Explicit (e.g. January 25, 2010), Implicit (e.g. New year's day 2009) or Relative(e.g. yesterday, next week etc.). At first, all temporal expressions must be tagged. The goals of socalled temporal taggers are the extraction of temporal expressions and the normalization of these expressions to some standard format as TIMEX 2 (consists of value, modifiers, normalized value of anchoring date or time, direction, set and comment) or TIMEX 3 (consists of offset, type and value). There are rule-based and machine learning-based approaches for the extraction of temporal expressions. But the normalization is done in a rule-based way. The research areas or trends of temporal information retrieval are Exploratory Search, Micro-blogging and Real-time Search, Temporal Summaries, Temporal Clustering, Temporal Querying, Temporal Question Answering, Temporal Similarity, Timelines and User Interfaces, Searching in Time, Web Archiving and Spatio-temporal Information Exploration.

The issue of P-time (Publication Time) detection and its application for page rank is addressed in [2]. An approach to extract P-time for a page with explicit P-time displayed on its body is proposed and then a method to infer P-time for a page without P-time is presented. Finally, a temporal sensitive page rank model using Ptime is discussed. Experiments demonstrate that these methods outperform the baseline methods significantly. If a page has explicit P-time in its HTML body then a domain and language independent machine learning method to extract the P-time is presented here. General linguistic and format information (Linguistic information, Position information, Format information, Tag information) are used to create 88 binary features for the machine learning model of Support Vector Machine (SVM) to identify the P-time. If a page does not have explicit P-time in its HTML body then it is inferred by the span of its P-time according to the link relation with its neighbors and then its exact P-time is inferred in terms of the text similarity between its content and those neighbors content who's P-time belongs to the span. An approach to rank pages considering their text content, temporal information (i.e. P-time in this paper), and page importance is proposed here. The hypothesis is that the text similarity of a page to a query does not change over time, while its importance changes over time.

The objective of [3] is to develop a retrieval system which can anticipate a user's likely temporal intent, considering recent or ongoing real-world events. Such a system should not only provide recent news when relevant, but also higher rank noontime stamped or even older documents which are temporally pertinent as they cover aspects related to recent event topics. Key challenges to be addressed in this work include: a suitable source and method for event detection and tracking, an intent-aware ranking approach and an evaluation methodology. For each intent, during ranking a measure of temporal intent pertinence is computed, thus higher ranking intents that refer to aspects related to recent events. Using Topic Detection and Tracking (TDT) techniques, Wikipedia article revision history and viewing counts can be mined for event-driven signals for many real-life topics, allowing the measurement of various temporal characteristics. An evaluation methodology based on query-log mining and crowd sourcing for on-going relevance assessment is proposed. Four research questions were proposed to investigate: Can search topics with recent event-related temporal intent be detected? Can the temporal sensitivity of a query topic (or, intent) be computed using historic and recent Wikipedia article revision history, and page view statistics? Given the temporal sensitivity of a query topic, can intent ranking be improved by incorporating temporal intent pertinence? If a query topic itself is not temporally sensitive, yet, an aspect is related to a recent event, can intent ranking be improved by incorporating temporal intent pertinence? [4] is based on explicit temporal query. Timeaware retrieval models exploit one of two time dimensions, namely, publication time or content time (temporal expressions mentioned in documents). The effectiveness for a temporal query (e.g. Illinois earthquake,1968) depends significantly on which time dimension is factored into ranking results. A machine learning approach is proposed to select the most suitable time-aware retrieval model for a given temporal query. This method uses three classes of features(Temporal KL-divergence, Clarity score, Retrieval scores) obtained from analyzing distributions over two time dimensions, a distribution over terms, and retrieval scores within top-k result documents. Temporal KL-divergence measures the difference between the distribution of publication time within a set of top-k result documents and their distribution in the overall document collection. The clarity score measures the KL-divergence between the distribution of terms within top-k results and their distribution in the overall document collection. Retrieval scores are measured by analyzing different features. It is demonstrated that selecting the right timeaware retrieval model can have a significant impact on the retrieval effectiveness of temporal queries. The novel machine learning approach is proposed here to do so automatically and demonstrated its effectiveness through extensive experiments.

In [5] the authors tried to develop a languageindependent model that tackles the temporal dimensions of a query and identifies its most relevant time periods. For this purpose, a temporal similarity measure capable of associating a relevant date(s) to a given query and filtering out irrelevant ones is proposed. This approach is based on the exploitation of temporal information from web content, particularly within the set of k-top retrieved web snippets returned in response to a query. It particularly focus on extracting years, which are a kind of temporal information that often appears in this type of collection. The methodology is evaluated using a set of real-world text temporal queries, which are clear concepts (i.e. queries which are nonambiguous in concept and temporal in their purpose). Experiments show that when compared to baseline methods, determining the most relevant dates relating to any given implicit temporal query can be improved with a new temporal similarity measure. This work presents a novel approach that aims to correctly tag the temporal expressions found in the documents, based on their relevance to the query and to properly tag implicit temporal queries with relevant years. This method is not based on metadata or query-logs, but on the exploitation of temporal information from the text itself. The proposal of this paper is : proposing a novel second-order similarity measure to assess the temporal similarity between a query and a date based on a content-based language-independent approach; exhaustively evaluating this measure on a real-world dataset and demonstrate extensive improvements when compared to state-of-the-art techniques; publicly providing a set of queries and ground-truth results to the research community.

[6] is based on implicitly year qualified query. Rather than solving the general problem of automatically determining user intent, this paper focuses on queries that have a temporally dependent intent. Temporally dependent queries are queries for which the best search results change with time. The search results for these queries should reflect the freshest, most current results. The algorithm relies only on having access to a query log with frequency information. It mines temporal patterns directly from query logs and do not make use of query frequency information or document timestamps. The foundations of the mining algorithm are built upon the assumptions: implicitly year qualified queries are strongly associated with several different years, and implicitly year qualified queries are associated with years more than they are associated with non-years. The mining algorithm takes a query as implicitly year qualified if it is qualified by at least two unique years. Even though a query is identified as implicitly year qualified does not necessarily mean that the query should always be treated as temporal in nature. This algorithm also finds these temporal ambiguities and checks if a query is always qualified with a year or not.

Freshness of web links is important to linkbased ranking algorithms. Old pages have more time to attract in-links, but may contain stale information. A single web snapshot is unable to detect sudden changes which might indicate link spam and further smooth or neutralize the undesirable influence automatically. In [7], an probabilistic algorithm is proposed to estimate web page authority by considering two temporal aspects. First, to avoid old pages from dominating the authority scores, to keep track of web freshness over time from two perspectives: how fresh the page content is, referred to as page freshness; and how much other pages care about the target page, referred as in-link freshness. To achieve this, web authors' maintenance activities on page content are mined. Each activity is associated with the time at which it occurs and temporal profiles for both pages and links are built. A random walk model is exploited to estimate the two predefined freshness measures. Multiple web snapshots at distinct time points are used, instead of a single snapshot. To make the link graph more stable, multiple web snapshots are connected by propagating authority flows among them, and so smooth the impact of sudden changes to particular snapshots on web page authority estimation. Several proximity-based density kernel functions are exploited to model such propagation. Combining web freshness measures, a semi-Markov process is utilized to model a web surfer's behavior in selecting and browsing web pages. The contributions of this work are : Quantify web freshness from authors' maintenance activities on web content over time, from the perspectives of page freshness and in-link freshness; Incorporate web freshness into authority propagation to favor fresh pages; Explore a series of proximity-based density kernel functions to model authority propagation among web snapshots; Conduct experiments on a real-world archival web data set and show the superiority of our approach on ranking performance in terms of both relevance and freshness.

4 III.

5 Proposed Algorithm a) Definitions

Explicit publication time of a web page refers to the time mentioned in the HTML body of a web page.

Inlink means the reference or link to the web page from other web pages.

Outlink refers to the reference or link from the web page to other web pages.

6 b) Algorithm

Step 1. Find publication time Using exact / explicit publication time or Time span detected from inlink and outlink and verified by matching.

Step 2. Extract the temporal concentration of the content and match it with publication time to make sure that the publication time is relevant or correct.

Step 3. Find the time span using all the documents retrieved.

Step 4. Build a binary tree using that time span.

Step 5. Show nodes from leaf to root.

IV.

7 Conclusion and Future Work

In the existing papers, we can see that some works are done for explicit temporal queries and some are done for implicit temporal queries. The works done for explicit temporal queries use both publication time and content time. They are easy to implement because the temporal information is given by the user as a part of the query. For implicit temporal query, the main challenge is to find out that the query indicates a specific time period. After finding out which queries have temporal intent, the rest of the work is done like explicit temporal query.

8 Global Journal of C omp uter S cience and T echnology

Volume XV Issue V Version I Year ( ) H [1] [2]

[1] © 2015 Global Journals Inc. (US) 1

[2] © 2015 Global Journals Inc. (US)

6

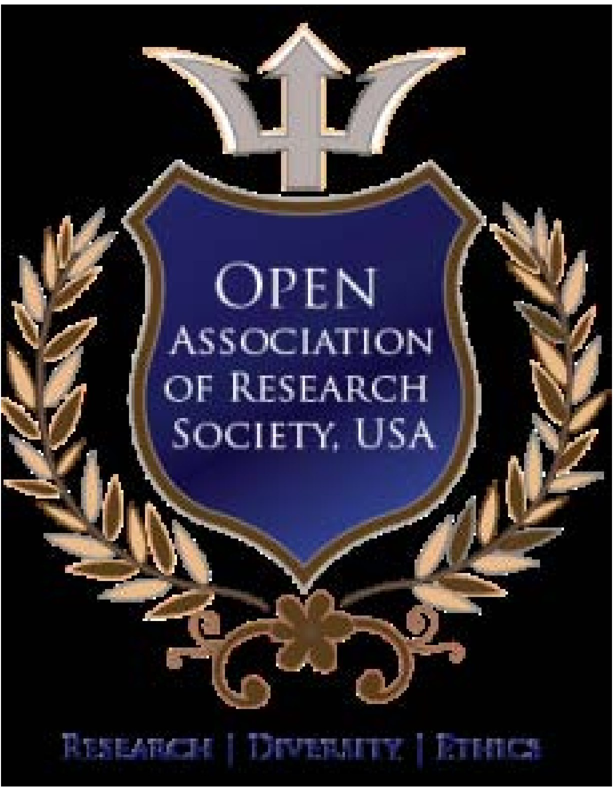

Figure 1: 6 Global

[Alonso et al. (2011)] O Alonso , J Strötgen , R B Yates , M Gertz . *Temporal Information Retrieval: Challenges and Opportunities*, (Hyderabad, India) 2011. March 28, 2011. TWAW.

[Dai and Davison ()] N Dai , B D Davison . *Freshness Matters: In Flowers, Food, and Web Authority, SIGIR'10*, (Geneva, Switzerland) July 19-23, 2010.

[Campos et al. ()] 'Enriching Temporal Query Understanding through Date Identification: How to Tag Implicit Temporal Queries?'. R Campos , G Dias , A M Jorge , C Nunes . *Temp Web '12*, (Lyon, France) Apr 16-17 2012.

[Metzler et al. ()] *Improving Search Relevance for Implicitly Temporal Queries, SIGIR'09*, D Metzler , R Jones , F Peng , R Zhang . July 19-23, 2009. Boston, Massachusetts, USA.

[Kanhabua et al. ()] *Learning to Select a Time-aware Retrieval Model, SIGIR'12*, N Kanhabua , K Berberich , K Nørvåg . August 12-16, 2012. Portland, Oregon, USA.

[Whiting ()] *The Essence of Time: Considering Temporal Relevance as an Intent-Aware Ranking Problem, SIGIR'12*, S Whiting . August 12-16, 2012. Portland, Oregon, USA.

[Chen et al. ()] *Web Page Publication Time Detection and its Application for Page Rank*, Z Chen , J Ma , C Cui , H Rui , S Huang . July 19-23, 2010. Geneva, Switzerland.