

#TAIA2012

Fernando Diaz
Microsoft Research
fdiaz@microsoft.com

Susan Dumais
Microsoft Research
sdumais@microsoft.com

Kira Radinsky
SalesPredict
kirar@cs.technion.ac.il

Maarten de Rijke
University of Amsterdam
derijke@uva.nl

Milad Shokouhi
Microsoft Research
milads@microsoft.com

Abstract

The SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)¹ was held in Portland on Thursday, 16 August 2012. The workshop brought together about 50 researchers from academia and industry for a full-day programme on time-sensitive information access that involved three keynotes and nine paper presentations.

1 Introduction

Web content increasingly reflects the current state of the physical and social world, manifested both in traditional news media sources along with user-generated publishing sites such as Twitter, Foursquare, and Facebook. At the same time, web searching increasingly reflects problems grounded in the real world.

As a result of this blending of the web with the real world, we observe that the web, both in its composition and use, has incorporated many of the dynamics of the real world. Many of the problems associated with searching dynamic collections are not well understood, such as defining time-sensitive relevance, understanding user query behavior over time and understanding why certain web content changes.

We aimed to bring together practitioners and researchers to discuss their recent breakthroughs and the challenges with addressing time-aware search, both from the algorithmic and the architectural perspectives.

The SIGIR 2012 Workshop on Time-Sensitive Information Access consisted of three keynote addresses, nine research presentations, and a breakout session.

¹<http://research.microsoft.com/en-us/people/milads/taia2012keynotes.aspx>

2 Keynotes

2.1 Neel Sundaresan: Relativity in Commerce Search

Abstract Transience and specificity (or lack thereof) introduce interesting challenges in commerce search. I will talk about introducing the “time dimension” into commerce search and the benefits of introducing this dimension. By mining user and inventory data at scale I show how the search engine can take advantage of this new dimension to bridge the gap between demand and supply. I will give examples from eBay, a well-known online marketplace and show, with examples, how useful and novel commerce experiences and applications can be afforded by understanding data at scale.

2.2 Rosie Jones: Time Sensitive Search and Shopping

Abstract Marketers view user purchase behavior as moving through a purchasing funnel over time—moving from phases such as awareness of a brand, to purchase intent, to the actual purchase. Users also move through different shopping phases during long-lasting tasks, such as shopping for goods and services while planning a wedding, or purchasing a new home then buying furniture. In this talk I discuss how historical shopping and search behavior can be used as features in a machine learned model for predicting future in-market shoppers for advertising. I will also discuss traditional time-sensitive marketing features such as frequency and recency, and how shopping cycles vary around holidays throughout the year.

2.3 Eytan Adar: Challenges for Temporal Information Interfaces

Abstract Growing interest in the temporal dynamics of information has led to a number of results and systems that have direct applicability to backend services (crawling, search, extraction, etc.). However, the added dimension of time has proven difficult when creating end-user interfaces. Many of the interfaces that exist today are simply insufficient to describe subtle changes in text for many documents, summarize large temporally-variant collections, and otherwise handle dynamic, and potentially ephemeral, datasets. The issue is not simply one of creating novel interaction patterns and understanding new use cases, but providing the back end architectures that reliably preform in the ways users have been trained to expect with current snapshot based systems that work on largely static collections. In my talk I will describe ongoing work, opportunities, and challenges in bringing temporal data to end users.²

3 Presentations

Research presentations covered a variety of subjects, grouped loosely into the topics of search, events, and interfaces. All papers and presentation slides are available for download at: <http://research.microsoft.com/en-us/people/milads/taia2012-accepted.aspx>

²More details about our keynote speakers can be found at: <http://bit.ly/RPfIGR>

3.1 Temporal Search

If a topic concerns a late-breaking news story or anything other issue with dynamic information, the relevance of documents for that topic will likely change as a function of time. Miles Efron began the session on temporal search by presenting recent work on modeling a document relevance using survival analysis. Efron's position is that the rate at which document relevance decays with time depends critically on the particular query. Some queries can be ranked purely by traditional relevance estimates while temporally-sensitive queries require the incorporate of recency information. As a result, he proposed a model that uses negative pseudo-relevance feedback to estimate the decay of relevance as a function of document age. His preliminary results demonstrated the effectiveness of this approach on TREC microblog data.

Anne Schuth presented evidence of the temporal sensitivity of answers in community-based question answering systems. In such systems, archived answers can be used to satisfy repetitions of previously asked questions. Schuth considers answers whose relevance does not change with time as *sustainable*. Schuth presented preliminary analysis of the sustainability of questions in a Yahoo! Answers corpus, though results appeared to be inconclusive.

Omar Alonso presented a position paper on integrating temporal characteristics into social search. Alonso proposed a taxonomy of temporal interest for an individual. Temporally sensitive topics which can be recent (i.e. new interests), ongoing (i.e. persistent interests), seasonal (i.e. interests which vary cyclically), past (i.e. interests which have faded), or random (i.e. interests which were ephemeral). Alonso proposed that the social dimension can be considered orthogonal to the temporal dimension. The social dimension ranges from personal interests to social interests to global interests. Each of these social levels can have temporal interests as defined by the taxonomy. If possible to classify a user's search session into one of these socio-temporal classes, Alonso argues that web search ranking can be improved.

3.2 Events on the web

Events refer to temporally acute topics in a stream. These can range in popular from large media events to relatively minor, highly personal events. Craig Macdonald presented an effort to reduce the amount of noise in event detection from Twitter by using page view information from Wikipedia. Specifically, the authors used high page view information as evidence supporting a tweet's discussion of a new event. The Wikipedia information was found to be especially useful for events with middling popularity. However, the authors also noted that events in the Wikipedia page views tended to lag Twitter, suggesting that the usefulness in a production setting may be limited.

In the spirit of combining evidence from multiple sources, Alonso presented work on automatically creating event timelines by synthesizing evidence from Wikipedia page views, query logs, and Twitter. Events in the timeline were based on Twitter hashtags whose time series were correlated with those found in related Wikipedia page views and query logs.

While many event detection and understanding systems leverage query and document volume, Nattiya Kanhabua advocated analyzing document content to extract events. Kanhabua proposed a model which incorporated evidence scoping from the sentence to document to corpus levels. When evaluated on a corpus of health-related reports, the model was effective at predicting the time of various disease outbreaks.

3.3 Temporal interfaces and activities

Daan Odijk presented their ongoing work on a coordinated time-aware exploratory search approach and presented a case study on a prototype system. He argued that using their approach, a user is able to gain a deeper understanding of the relevant parts of the collection, and this is particularly useful on historical archives that span over centuries. They proposed an interactive search system with an interface that combines faceted search techniques and interactive visualizations. These visualizations use methods illustrate the volume and correlation of documents and terms over time, and are presented alongside the search results and facets.

The public Wikipedia page view logs appeared in several presentations as a valuable source of auxiliary information. Researchers at the University of Amsterdam presented OpenGeist, a public web service intended to make Wikipedia page view data easier to analyze. The service allows users to request raw time series data associated with a Wikipedia page. In addition, the service provides the ability to return summary statistics of a time series and to group related pages.

Maarten de Rijke gave an overview of his joint work with Wouter Weerkamp on activity prediction in twitter. For activity prediction one need to decide for which future timeframe the predictions should be made? In their work the focus is on *tonight*. Preliminary results based on simple keyword matching suggests that effective activity prediction is feasible. The authors identified seven directions for future research namely: activity classification, summarization, term extraction, time indication, tweet segmentation, evaluation, and combining multiple sources.

4 Discussion

Throughout the workshop, we noted several recurring themes for further discussion: assembling test collections with temporally sensitive tasks, dealing with multiple temporal data streams, and intent modeling.

The discussion of test collections requires first defining a temporally sensitive task. One option would be to evaluate an *ad hoc* ranking for a subset of queries with temporally sensitive relevance. These queries could either be extracted from public ‘trending queries’ or crafted by assessors. Some raised logistical issues with assessing temporally sensitive queries, known to be a problem encountered when assessing relevance for GeoCLEF. Another option would be to revisit the Topic Detection and Tracking (TDT) tasks with modern corpora and more time-sensitive metrics. There was a suggestion for defining a task which tracked how a new story evolved; this would be similar to TREC’s Knowledge Base Acceleration Track with a focus on temporal information. Finally, another novel task proposed automatically creating a calendar from microblog data. Any task would require a set of timestamped documents. The most recent ClueWeb crawl includes fresh URLs gathered from the Twitter Gardenhose stream. Other sources of timestamped documents include Twitter (used in the TREC microblog track), Wikipedia edits, and Reuters.

Several research presentations noted that using multiple data streams improved performance. In addition to Wikipedia page views and microblog data, participants suggested even more streams be considered, including data from finance, the weather, and even EEG. More quantitative data can be used for evaluation, for example if considering an information extraction task. In some cases, different streams represent the consumption and production

in information systems (i.e. queries and documents) and is not well understood.

5 Summary

The workshop successfully attracted researchers interested in temporally sensitive search tasks. Both prior and ongoing work were discussed and there was an interest in holding another workshop at future conferences.