DiversiWeb 2011

Elena Simperl
KIT
Karlsruhe, Germany
elena.simperl@kit.edu

Devika P. Madalli
Indian Statistical Institute
Bangalore, India
devika@drtc.isibang.ac.in

Denny Vrandečić
KIT
Karlsruhe, Germany
denny.vrandecic@kit.edu

Enrique Alfonseca
Google
Zurich, Switzerland
ealfonseca@google.com

Abstract

DiversiWeb 2011, the First International Workshop on Knowledge Diversity on the Web, co-located with WWW2011 in Hyderabad, India, provided an interdisciplinary forum for researchers and practitioners to present and discuss their ideas related to the challenges posed by diversity on the Web. We addressed a wide array of interdisciplinary questions, which need to be tackled in order to preserve the fragile balance between a world that is continually converging and growing together, the rich diversity of the global society, and the dangers of fragmentation and splintering.

1 Introduction

Almost 20 years after its introduction, the Web provides a platform for the publication, use and exchange of information, at planetary scale, on virtually every topic, and representing an amazing diversity of opinions, viewpoints, mindsets and backgrounds. The success of the Web can be attributed to several factors, most notably to its principled scalable design, but also to a number of subsequent ICT developments such as smart user-generated content, mobile devices, and most recently cloud computing.

The first two of these have dramatically lowered the last barriers of entry when it comes to producing and consuming information online, leading to an unprecedented growth and mass collaboration. They are responsible for hundreds of millions of users all over the globe creating high-quality encyclopedias, publishing Terabytes of multimedia content, contributing to world-class software, and lively taking part in defining the agenda of many aspects of our society by raising their voices, and publicly expressing and sharing their ideas, viewpoints, and resources.

The other side of the coin in this unique success story is, nevertheless, the great challenges associated with managing the sheer amounts of information continuously being published online, whilst allowing for purposeful use, and leveraging the diversity inherently unfolding through global-scale collaboration. In this context, diversity includes different opinions, sentiments, preferences, or worldviews that are reflected in the way information is expressed on the Web. These challenges are still to be solved at many levels, from the infrastructure to store and access the information, through the methods and techniques to make sense out of it, to the paradigms underlying the processes of Web-based information provision and consumption.

As an example, when searching for blog posts, state-of-the-art technology – be that popularity-based algorithms, recommendation engines or collaborative filters – tends to return either the most popular posts, or those which correspond with a personal profile and therefore with the known opinions and tastes of the reader. Alternative points of view, and new unexpected content, are not taken into account as they are not highly ranked, and posts expressing different opinions are sometimes even discarded.

This behavior has particularly negative consequences when dealing with information that is expected and intended to be subject to diverse opinion – as is the case with news reports, ratings of products or media content, customer reviews, or any other type of subjective assessment. The same negative effects apply in a community-driven environment that is designed for collaboration – the most obvious example here being Wikipedia and the blogosphere. The information diversity exposed in such an environment, impressive both with respect to scale and the richness of opinions and viewpoints expressed, cannot be handled without adequate computer support in an economically feasible manner. In the long run, maintaining the current state-of-affairs will change the ways and the extent to which people are informed (or not) on a particular topic, tremendously influencing how they look into that topic, what they find about it and what they think about it.

On top of all this, it is meanwhile acknowledged that the current state of affairs hampers true collaboration. For example, Wikipedia is a tremendous success, but it is also a largely meritocratic system with a decreasing number of active contributors, whereas the blogosphere has to deal with the limited attention of the blog authors. What is needed are novel concepts, methods and tools that allow humans and machines to leverage the huge amounts of information created by a community, based on interaction models that support expressing, communicating and reasoning about divergent models simultaneously. This would not only enhance true collaboration, but would also significantly improve various aspects of the information management life cycle, thus addressing information overload in sectors which rely on opinions-driven information sources and mass participation – news, ratings, reviews, and social and information sharing portals of any kind.

2 Goals of the workshop

The overall aim of this workshop was to provide an interdisciplinary forum for researchers and practitioners to present and discuss their ideas related to the challenges posed by diversity on the Web. We addressed a wide array of interdisciplinary questions, which need to be tackled in order to preserve the fragile balance between a world that is continually converging and growing together, the rich diversity of the global society, and the dangers of fragmentation and splintering. This includes but is not limited to questions such as *'How to model diversity?'*, *'How to discover bias and opinion in blog posts, tweets, forum items, wiki edits, etc.?'*, *'How to rank, aggregate, summarize, and exploit information in a diversity-aware manner?'*, *'What* *are the applications of diversity-rich information sources?', 'How can we use diversity as an asset instead of regarding it as a barrier?'.*

3 Topics

The list of topics mentioned in the call for papers was:

- Analyze the capabilities of current information management models, algorithms and technologies to leverage knowledge diversity,
- Extend existing models, methods, techniques and tools to accommodate the requirements arising from paying a proper account to diversity-expressed information sources and communication and collaboration environments characterized by a rich variety of opinions and viewpoints.
- Discuss the foundations of knowledge diversity on the Web and propose alternative paradigms,
- Propose novel evaluation strategies, methods and techniques to assess the impact of diversity-minded information management.

Topics of the workshop include, but are not limited to:

- Risks and advantages of diversity and diversification on the Web
- Facets of knowledge diversity and conceptual and formal models for representing and understanding diversity
- Discovery and mining of corpora for diversity-related information
- Use of Natural Language Processing techniques for diversity mining
- Classifying Web 2.0 content items such as blog posts, videos, tweets, and wiki-edits by their biases
- Usage and benefits of diversity in the corporate context, e.g. in order to understand feedback and communication with the customer
- Enabling or improving communication and collaboration over barriers induced by diversity
- Extensions to Web applications taking diversity into account
- Exposing and explaining diversity to end users
- User experiences avoiding the radicalization of groups by exposing them to alternatives
- User interfaces allowing the explicit annotation of content with diversity markers
- Studies on the acceptance by end-users of diversified applications.

4 Papers

We selected the following six papers from the submissions, based on the recommendations of the DiversiWeb programme committee.[1]

[1] `http://render-project.eu/diversiweb-2011/`

- **Towards a Knowledge Diversity Model** by Rakebul Hasan, Katharina Siorpaes, Reto Krummenacher, and Fabian Flöck. *Abstract:* The Web is an unprecedented enabler for publishing, using and exchanging information at global scale. Virtually any topic is covered by an amazing diversity of opinions, viewpoints, mind sets and backgrounds. The research project RENDER works on methods and techniques to leverage diversity as a crucial source of innovation and creativity, and designs novel algorithms that exploits diversity for ranking, aggregating and presenting Web content. Essential in this respect is a knowledge model that makes accessible – cognitively to human users as well as computationally to the machine – the diversity in content. In this paper, we present a glossary of relevant terms that serves as baseline to the specification of the Knowledge Diversity Model.
- **Expressing Opinion Diversity** by Andreea Bizău, Delia Rusu, and Dunja Mladenić. *Abstract:* The focus of this paper is describing a natural language processing methodology for identifying opinion diversity expressed within text. We achieve this by building a domain-driven opinion vocabulary, in order to be able to identify domain specific words and expressions. As a use case scenario, we consider Twitter comments related to movies, and try to capture opinion diversity by employing an opinion vocabulary, which we generate based on a corpus of IMDb movie reviews.
- **Scalable Detection of Sentiment-Based Contradictions** by Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. *Abstract:* The analysis of user opinions expressed on the Web is becoming increasingly relevant to a variety of applications. It allows us to track the evolution of opinions or discussions in the blogosphere, or perform product surveys. The aggregation of sentiments and analysis of contradictions is another important application, which becomes effective since we are able to capture the diversity in sentiments on different topics with more precision and on a large scale. Though, there is still a need for a scalable way of sentiment aggregation with respect to the time dimension, which preserves enough information to capture contradictions. In this paper, we are focusing on the problem of finding sentiment-based contradictions at a large scale. First, we define two types of contradictions, depending on the distributions of opposite sentiments over time. Second, we introduce a novel measure of contradiction based on the mean value and the variance of sentiments among different texts. Third, we propose a scalable method for identifying both types of contradictions at different time scales. We evaluate the performance of our method using synthetic and real-world datasets, as well as a user-study. The experiments demonstrate the effectiveness of the proposed method in capturing contradictions in a scalable manner.
- **Faceted Approach To Diverse Query Processing** by Alessandro Agostini, Devika P. Madalli, and A. R. D. Prasad. *Abstract:* This paper presents a formal framework for implementing a query refinement method. The method uses general principles of facet analysis. Two key notions are advanced and discussed: diversity and focus. Diversity refers to the information needs of a querying user; it is captured by the notion of facet. A focus refers to how diversity is captured from the documents as organized by the user; it provides a kind of context to the user query. The method is situated within the formal framework of the smallest propositionally closed description logic $\mathcal{ALC}$, thereby betting that $\mathcal{ALC}$ provides us with a suitable SAT solver to implement a facet engine, which is the main component of our method.
- **Approximate subgraph matching for detection of topic variations** by Mitja

Trampuš, and Dunja Mladenić. *Abstract:* The paper presents an approach to detection of topic variations based on approximate graph matching. Text items are represented as semantic graphs and approximately matched based on a taxonomy of node and edge labels. Best-matching subgraphs are used as a template against which to align and compare the articles. The proposed approach is applied on news stories using WordNet as the predefined taxonomy. Illustrative experiments on real-world data show that the approach is promising.

- **Mining Diverse Views from Related Articles** by Ravali Pochampally, and Kamalakar Karlapalem. *Abstract:* The world wide web allows for diverse articles to be available on a news event, product or any topic. It is not impossible to find a few hundred articles that discuss a specific topic thus making it difficult for a user to quickly process the information. Summarization condenses huge volume of information related to a topic but does not provide a delineation of the issues pertaining to it. We want to extract the diverse issues pertaining to a topic by mining views from a collection of articles related to it. A view is a set of sentences, related in content, that address an issue relevant to a topic. We present a framework for extraction and ranking of views and have conducted experiments to evaluate the framework.

5 Video and paper proceedings

The proceedings of the workshop are available from the workshop website.[2] Video recordings of all talks as well as the slides of the talks are expected to be available by mid 2011 thanks to courtesy of VideoLectures.net.[3]

6 Acknowledgments

The workshop was partially funded by EU IST FP7 projects RENDER[4] and Living Knowledge.[5] We thank the program committee for the indispensable help in the selection process for the workshop program, and VideoLectures.net for recording the workshop.

[2] http://render-project.eu/diversiweb2011/

[3] http://videolectures.net

[4] http://render-project.eu

[5] http://livingknowledge-project.eu/