Diversity and Novelty on the Web: Search, Recommendation, and Data Streaming Aspects

Rodrygo L. T. Santos
Univ. Federal de Minas Gerais
Belo Horizonte, MG, Brazil
rodrygo@dcc.ufmg.br

Pablo Castells
Univ. Autónoma de Madrid
Madrid, Spain
pablo.castells@uam.es

Ismail Sengor Altingovde
Middle East Technical University
Ankara, Turkey
altingovde@ceng.metu.edu.tr

Fazli Can
Bilkent University
Ankara, Turkey
canf@cs.bilkent.edu.tr

ABSTRACT

This tutorial aims to provide a unifying account of current research on diversity and novelty in different web information systems. In particular, the tutorial will cover the motivations, as well as the most established approaches for producing and evaluating diverse results in search engines, recommender systems, and data streams, all within the context of the World Wide Web. By contrasting the state-of-the-art in these multiple domains, this tutorial aims to derive a common understanding of the diversification problem and the existing solutions, their commonalities and differences, as a means to foster new research directions.

1. INTRODUCTION

Diversification has been traditionally studied in the context of search, notably web search [3]. In particular, web search queries are typically short and often ambiguous. Handling such queries involves identifying relevant search results for users with possibly rather different information needs for the same query, while incurring minimum redundancy in the resulting ranking. *Implicit diversification approaches* for search results presume no prior information about the possible different user needs (a.k.a. aspects, interpretations, intents, etc.) of a given query, so the diversification is carried out by somehow inferring these needs and/or computing the diversity from the set of documents initially retrieved for the query. Instead of implicitly assuming that similar documents will cover similar needs of the query, the broad topic underlying an ambiguous or underspecified query can be usually decomposed into its constituent sub-topics. This is precisely the intuition behind *explicit diversification approaches* (e.g., [2]). These approaches have exploited query properties such as the query's possible categories or its reformulations, mined from the query logs of web search engines.

Diversity, along with novelty, has also been recognized and researched in the area of *recommender systems* as a key dimension of the system output utility, with different angles but obvious—yet to some degree unexplored—connections to diversity in web search [4]. On the one hand, the recommender systems community has grown increasingly aware that the accuracy in matching user preferences alone is not enough for users to find value in recommendations. In many, if not most scenarios, the whole point of recommendation is inherently linked to a notion of discovery, as recommendation makes most sense when it exposes the user to a relevant experience she would not have found, or thought of by herself, whereby obvious, however accurate recommendations are generally of little use. On another axis, recommendations procure a richer experience if the retrieved items are not too similar and redundant to each other, in order to provide the user with a wider array of choice. Finally, sales diversity can help enhance businesses as well, as a strategy to leverage revenues from market niches.

Stream processing applications aim to analyze continuous data streams and generate output streams for information consumers. It is possible to see stream processing applications in several problem domains such as financial markets, health care, traffic monitoring, intelligence applications, computational social sciences, etc. Diversification and novelty detection in continuous document data or *document data streams* (DDS) are extensively used in publish/subscribe systems and they are used for information filtering, information aggregation, trend detection, sentence extraction-based summarization, etc. One typical application of novelty detection in DDS is new event detection also known as first story detection [1]. Diversification in stream processing aims to bring existing or possible future objects of interests to the users' attention, if needed in a customized fashion. Like in other novelty and diversification applications, the ultimate purpose is to prevent overwhelming users with unnecessary/known data.

2. TUTORIAL OVERVIEW

This tutorial aims to provide a unifying account of current research on diversity and novelty in multiple Web information systems. In particular, we will cover the motivations, as well as the most established approaches for producing and evaluating diverse results in the context of search engines,

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741988.

recommender systems, and data streams. By contrasting the state-of the-art in these multiple domains, this tutorial aims to derive a common understanding of the diversification problem and the existing solutions, their commonalities and differences, as a means to foster new research directions. In particular, the tutorial attendees will:

- understand the importance and complexities of achieving diversity/novelty for various Web domains;
- learn the state-of-the-art approaches for promoting diversity and novelty in web search, recommender systems, and data streams;
- learn the fundamental evaluation metrics and have an overview of past and current evaluation campaigns;
- get an overview of other related application areas that include query suggestions, query ambiguity detection, and aggregated search;
- obtain a unifying view of the topic by exploring the similarities and differences between the methods employed in different domains.

3. TUTORIAL OUTLINE

1. Practical and Theoretical Background
2. Diversity and Novelty in Web Search
   - *Implicit and Explicit Diversification*
   - *Advanced Topics in Web Search Diversification*
   - *Evaluation*
3. Diversity and Novelty in Recommender Systems
   - *Motivation and Notions*
   - *Novelty and Diversity Enhancement*
   - *Evaluation*
4. Diversity and Novelty in Data Streams
   - *Document-level Novelty*
   - *Novelty and Diversification of Document Streams*
   - *Evaluation*

4. PRESENTERS' BIOGRAPHY

Rodrygo Santos is an assistant professor at the Federal University of Minas Gerais (UFMG), Brazil. He holds BSc (2005) and MSc (2007) degrees from UFMG, and a PhD (2013) from the Univ. of Glasgow, UK. His research interests encompass large-scale search and recommendation in various domains, including the Web, social media, and enterprises. In addition, he is a leading expert in search result diversification, with a PhD thesis, a book, and over 20 research papers in the field over the past five years. As a result of his work, he has contributed to the open-source Terrier Information Retrieval platform, and published research papers in several major academic venues. In particular, he is a regular speaker and program committee member of several premiere conferences in IR, including SIGIR, CIKM, ECIR, ICTIR, and SPIRE, and a member of the editorial board of Foundations and Trends in Information Retrieval.

Pablo Castells is an associate professor at the Autónoma University of Madrid. His research experience is focused in the areas of information retrieval, recommender systems and personalization. He has led or participated in several national and international projects and has co-authored over 70 journal and conference publications in the aforementioned areas, serving regularly as a reviewer and scientific committee member for international journals and conferences in or related to IR. In recent years his research has focused on diversity, novelty and evaluation in IR and recommender systems, with publications in venues such as RecSys and SIGIR. He has been guest editor of three journal special issues on IR personalization, recommender systems, novelty and diversity in IP&M, ACM TIST (in preparation), and SIVP, respectively. Among other venues, he co-organized a workshop at each of the last four ACM RecSys conferences, around the topics of diversity and evaluation for recommender systems, and a workshop on personalized IR evaluation at SIGIR 2013. He was recently involved in the organization of the ESSIR 2013 summer school.

Ismail Sengor Altingovde is an assistant professor in the Computer Engineering Dept. of Middle East Technical Univ. (Turkey). He has received his BSc, MSc and PhD degrees, all in Computer Engineering, from Bilkent University (Turkey) in 1999, 2001 and 2009, respectively. Before joining METU, he worked as a postdoctoral researcher at Bilkent and L3S Research Center in Germany. His research interests include web IR, with a particular focus on search efficiency, and social web; as well as the effectiveness and efficiency of search results on the Web and social platforms. He has published over 40 papers in prestigious journals (including ACM TODS, ACM TOIS, ACM TWEB, JASIST and IP&M) and conferences (including SIGIR, VLDB, and CIKM). Most recently, he lectured at the Russian Summer School in IR (2012) and co-organized the LSDS-IR Workshop at WSDM 2013 and 2014. He is one of the recipients of Yahoo! Faculty Research and Engagement Program (FREP) award in 2013.

Fazli Can is a professor of Comp. Eng. at Bilkent Univ. in Ankara, Turkey. Before Bilkent he was a tenured professor in the CS Dept. at Miami Univ., Oxford, OH where he taught between 1986 and 2005. He was one of the two co-editors of ACM SIGIR Forum (1995-2002). He has extensively published in IR, data mining, database, computational linguistics, multimedia conferences and journals such as IP&M, JASIST, ACM TOIS, ACM TODS, and Multimedia Tools and Applications. He has served on several PCs and he was one of the general co-chairs of the IEEE/ACM ASONAM 2012 Conference. His recent works on topic tracking and novelty detection have been published in JASIST. He has recently co-edited a book titled *State of the Art Applications of Social Network Analysis* published by Springer in 2014.

Acknowledgments. I. S. Altingovde and F. Can are partially supported by METU BAP-08-11-2013-055 project and The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant no. 113E249, respectively.

5. REFERENCES

[1] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar. New event detection and topic tracking in Turkish. *JASIST*, 61(4):802–819, 2010.

[2] A. M. Ozdemiray and I. S. Altingovde. Explicit search result diversification using score and rank aggregation methods. *JASIST*. In press. http://dx.doi.org/10.1002/asi.23259

[3] R. L. T. Santos, C. Macdonald, and I. Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, 2015.

[4] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proc. of ACM RecSys*, pages 109–116, 2011.