Integrating and Ranking Aggregated Content on the Web

The Theory and Practice of Aggregated Search and Whole-Page Composition

Jaime Arguello
School of Information and
Library Science
University of North Carolina
Chapel Hill, NC
jarguello@unc.edu

Fernando Diaz*
Yahoo! Labs New York
111 W 41st Street
17th Floor
New York, NY
diazf@yahoo-inc.com

Milad Shokouhi
Microsoft Research
Cambridge, United Kingdom
milads@microsoft.com

ABSTRACT

Commercial information access providers increasingly incorporate content from a large number of specialized services created for particular information-seeking tasks. For example, an aggregated web search page may include results from image databases and news collections in addition to the traditional web search results; a news provider may dynamically arrange related articles, photos, comments, or videos on a given article page. These auxiliary services, known as *verticals*, include search engines that focus on a particular domain (e.g., news, travel, or sports), search engines that focus on a particular type of media (e.g., images, video, or audio), and APIs to highly-targeted information (e.g., weather forecasts, map directions, or stock prices). The goal of *content aggregation* is to provide integrated access to all verticals within a single information context. Although content aggregation is related to classic work in distributed information retrieval, it has unique signals, techniques, and evaluation methods in the context of the web and other production information access systems.

In this tutorial, we present the core problems associated with content aggregation, which include: sources of predictive evidence, sources of training data, relevance modeling, and evaluation. While much of the aggregation literature is in the context of web search, we also present material related to aggregation more generally. Furthermore, we present material from both academic and commercial perspectives and review solutions developed in both environments, which provides a holistic view for researchers and a set of tools for different types of practitioners.

1. RELATED TUTORIALS

Previous workshops and tutorials have covered somewhat similar material. The SIGIR 2008 Workshop on Aggregated Search was the the first forum to discuss matters related to aggregation [14]. However, at the time, the field was relatively new and preliminary experiments had not yet been conducted. The ECIR 2010 tutorial on distributed information retrieval covered federation in general. However, most of these techniques assumed homogeneous distributed search engines. The shift towards integrating search engines that focus on *different* types of media and serve *different* information seeking tasks require new techniques, new sources of predictive evidence, and new evaluation methods.

Two of the authors, Fernando Diaz and Milad Shokouhi, together with Mounia Lalmas, presented a related tutorial at SIGIR 2010. The proposed tutorial builds on this by incorporating significant new research published since this initial tutorial. This new research includes evaluation methodologies for experiments in academic [3] and production settings [11]. In addition, whereas the previous tutorial presented methods for federation in general, we focus specifically on the web setting, allowing a deeper discussion of nuances in this scenario, and also cover related areas such as whole-page composition that share similar challenges.

2. OVERVIEW

Broadly speaking, the goal of content aggregation is to combine information from different sources (i.e., different search engines, databases, or applications) into a single presentation. This tutorial takes a general approach to content aggregation, bringing together applications such as personalized news content aggregation [9, 8, 11, 10], context-aware business recommendation on mobile devices [21], and aggregated web search [4, 15, 6]. Within the context of personalized news content aggregation, given a particular user or user population, the system must predict which auxiliary content to present and where to present it relative to the core news story. Similarly, within the context of aggregated web search, given an information request, an aggregated web search system must predict which verticals to present and where to present them relative to the core web results.

Current methods for content aggregation build upon techniques developed to solve other types of problems. This tutorial provides a high-level overview of these related areas: information filtering [9, 8], distributed information retrieval [17], meta-search, peer-to-peer search [13], data-fusion, and topic detection and tracking [1].

A significant portion of the tutorial focuses on aggregated web search, which is typically decomposed into two subsequent tasks: predicting *which* verticals to present (*vertical selection*) and predicting *where* in the web results to present them (*vertical presentation*). State-of-the-art methods for vertical selection and presentation use machine learning to combine different types of features [4, 7, 2, 5, 6, 15, 19, 12]. A major goal of the tutorial is to motivate and describe these different types of features. Query features focus on properties of the query string itself, independent of any resource associated with a vertical [4, 15, 12]. Vertical corpus features focus on the similarity between the query and content from

*primary contact

the vertical [4, 18, 6]. Vertical query-log features focus on the similarity between the query and queries issued directly to the vertical by users [4, 6]. Vertical click-through features focus on clicks and skips on previous presentations of the vertical for the same query [15, 6, 19] or similar queries [7].

Relevance modeling refers to how a system can make vertical relevance predictions as a function of a set of features. In this tutorial, we review two different ways of modeling relevance. Off-line models are trained once using either vertical-relevance judgements [4, 12] or user-generated clicks and skips [15]. On-line methods can dynamically adjust their parameters in the presence of implicit user feedback. [6, 7].

Aggregated web search evaluation remains an open area of research. The tutorial provides a description of different types of evaluation, including batch-level evaluation [3, 4], user-study evaluation [20], and user-interaction-based evaluation [15]. We discuss the advantages and disadvantages of each type of approach and describe how each can be used to answer different questions.

The tutorial concludes with an overview of special topics in content aggregation such as modeling changes in user interest [16], domain-adaptation for vertical selection [5], and explore/exploit methods for harnessing user feedback [10]. We also discuss potential areas for new research.

The presentation slides for this tutorial are available at: `http://ils.unc.edu/~jarguell/www12_content_agg/`

3. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*, volume 12 of *The Information Retrieval Series*. Springer, New York, NY, USA, 2002.

[2] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *CIKM 2011*, pages 201–210. ACM, 2011.

[3] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *ECIR 2011*, pages 141–152. Springer-Verlag, 2011.

[4] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR 2009*, pages 315–322. ACM, 2009.

[5] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR 2010*, pages 691–698. ACM, 2010.

[6] F. Diaz. Integration of news content into web results. In *WSDM 2009*. ACM, 2009.

[7] F. Diaz and J. Arguello. Adaptation of offline vertical selection predictions in the presence of user feedback. In *SIGIR 2009*. ACM, 2009.

[8] A. Jennings and H. Higuchi. A personal news service based on a user model neural network. *IEICE Transactions on Information and Systems*, 75(2):192–209, 1992.

[9] T. Kamba, K. Bharat, and M. C. Albers. The krakatoa chronicle - an interactive, personalized, newspaper on the web. In *WWW 1995*, pages 159–170. ACM, 1995.

[10] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW 2010*, pages 661–670. ACM, 2010.

[11] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM 2011*, pages 297–306. ACM, 2011.

[12] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *SIGIR 2008*, pages 339–346. ACM, 2008.

[13] J. Lu. *Full-text federated search in peer-to-peer networks*. PhD thesis, Language Technologies Institute, Carnegie Mellon University, 2007.

[14] V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2):80–83, Nov. 2008.

[15] A. K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo. On composition of a federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. In *WSDM 2011*, pages 715–724. ACM, 2011.

[16] A. Saha and V. Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *WSDM 2012*, pages 693–702. ACM, 2012.

[17] M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[18] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *SIGIR 2003*, pages 298–305. ACM, 2003.

[19] Y. Song, N. Nguyen, L.-w. He, S. Imig, and R. Rounthwaite. Searchable web sites recommendation. In *WSDM 2011*, pages 405–414. ACM, 2011.

[20] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM 2010*, pages 519–528. ACM, 2010.

[21] J. Zhuang, T. Mei, S. C. Hoi, Y.-Q. Xu, and S. Li. When recommendation meets mobile: contextual and personalized recommendation on the go. In *UbiComp 2011*, pages 153–162. ACM, 2011.