Information Retrieval with Verbose Queries

Manish Gupta
Microsoft
gmanish@microsoft.com

Michael Bendersky
Google, Inc.
bemike@google.com

1. COVER SHEET

Proposed duration is a full-day tutorial. The current plan is to divide the tutorial into two main parts, each focusing on applications of the discussed techniques to verbose natural language queries.
1. Query reduction, reformulation and segmentation techniques.
2. Query concept weighting, expansion and learning-to-rank.

Contact Information

Manish Gupta (main contact):
Email: gmanish@microsoft.com
Address: 3B6014, Microsoft Building 3, Microsoft Campus, Gachibowli, Hyderabad-500032, India.

Michael Bendersky
Email: bemike@google.com
Address: Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043.

Intended Audience

Information retrieval with verbose natural language queries has gained a lot of interest in recent years both from the research community and the industry. Search with verbose queries is one of the key challenges for many of the current most advanced search platforms, including question answering systems (Watson or Wolfram Alpha), mobile personal assistants (Siri, Cortana and Google Now), and entity-based search engines (Facebook Graph Search or Knowledge Graph). Therefore, we believe that a tutorial on this topic at SIGIR is very timely and will attract a lot of interest from all conference participants.

Researchers in the field of analysis of search queries will benefit the most, as this tutorial will give them an exhaustive overview of the research in the direction of handling verbose web queries. We believe that the tutorial will give the newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more. Practitioners and people from the industry will clearly benefit from the discussions both from the methods perspective, as well from the point of view of applications where such mechanisms are starting to be applied.

After the tutorial, the audience will be able to appreciate and understand the following: (1) What are the interesting properties of complex natural language verbose queries; (2) Challenges in effective information retrieval with verbose queries; (3) State-of-the-art techniques for verbose query transformations that yield better expected search performance; (4) State-of-the-art ranking methods for verbose queries, including supervised learning-to-rank methods (5) What user/industry segments can be affected by better retrieval with verbose queries and what are the possible applications.

Pre-requisites: Introductory-level knowledge in information retrieval, query log mining, web mining, algorithms, natural language processing and machine learning.

Brief Biography

Manish Gupta (*Homepage Link*) is a Senior Applied Scientist at the Bing team in Microsoft India R&D Private Limited at Hyderabad, India. He is also an Adjunct Faculty at International Institute of Information Technology, Hyderabad. He received his Masters in Computer Science from IIT Bombay in 2007 and his Ph.D. from the University of Illinois at Urbana-Champaign in 2013. Before this, he worked for Yahoo! Bangalore for two years. His research interests are in the areas of web mining, data mining and information retrieval. He has published more than 30 research papers in referred journals and conferences, including WWW, SIGIR, ICDE, KDD, PKDD, SDM conferences. He has also co-authored a book on Outlier Detection for Temporal Data.

Manish has an extensive experience in offering tutorials at top conferences. Following is a list of recent tutorials he has offered.

- **Manish Gupta**, Rui Li, Kevin C. Chang. Towards a Social Media Analytics Platform: Event and Location Detection for Microblogs. *WWW 2014*.
- **Manish Gupta**, Jing Gao, Charu Aggarwal, Jiawei Han. Outlier Detection for Temporal Data. *CIKM 2013*.
- **Manish Gupta**, Jing Gao, Charu Aggarwal, Jiawei Han. Outlier Detection for Graph Data. *ASONAM 2013*.
- **Manish Gupta**, Jing Gao, Charu Aggarwal, Jiawei Han. Outlier Detection for Temporal Data. *SDM 2013*.

He also taught a full credit course on "Web Mining" at IIIT-Hyderabad, India in 2013 and in 2014. He also participated in the organization of CMU Winter School in 2014.

Michael Bendersky (*Homepage Link*) is a Senior Software Engineer at Google, where he works on organizing the world's information and making it universally accessible and useful. He received his Ph.D. from the University of Massachusetts Amherst in 2012. Michael published more than 20 research papers on information retrieval with verbose natural language queries. His paper "Discovering Key Concepts in Verbose Queries", published at SIGIR 2008, has been widely cited as one of the seminal works in this research area. Since then, his papers on query segmentation, query

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767877.

expansion and query representations for information retrieval appeared at top-tier academic conferences, including SIGIR, CIKM, WSDM, WWW, ACL and SIGKDD.

Michael co-organized a successful series of workshops on "Query Representation and Understanding" held at SIGIR 2010 and 2011. He also served as a publicity chair for the WSDM 2014 conference.

2. MOTIVATION

Recently, the focus of many novel search applications shifted from short keyword queries to verbose natural language queries. Examples include question answering systems and dialogue systems, voice search on mobile devices and entity search engines like Facebook's Graph Search or Google's Knowledge Graph. However the performance of textbook information retrieval techniques for such verbose queries is not as good as that for their shorter counterparts. Thus, effective handling of verbose queries has become a critical factor for adoption of information retrieval techniques in this new breed of search applications.

Over the past decade, the information retrieval community has deeply explored the problem of transforming natural language verbose queries using operations like reduction, weighting, expansion, reformulation and segmentation into more effective structural representations. However, thus far, there was not a coherent and organized tutorial on this topic. In this tutorial, we aim to put together various research pieces of the puzzle, provide a comprehensive and structured overview of various proposed methods, and also list various application scenarios where effective verbose query processing can make a significant difference.

3. OBJECTIVES

After the tutorial, the audience will be able to appreciate and understand the following: (1) What are the interesting properties of complex natural language verbose queries; (2) Challenges in effective information retrieval with verbose queries; (3) State-of-the-art techniques for verbose query transformations that yield better expected search performance; (4) State-of-the-art ranking methods for verbose queries, including supervised learning-to-rank methods (5) What user/industry segments can be affected by better retrieval with verbose queries and what are the possible applications.

4. RELEVANCE TO INFORMATION RETRIEVAL COMMUNITY

As can be seen from the long list of referenced papers, in this tutorial, we will organize related work done by the information retrieval community in the past decade, present it as a coherent story, and summarize the research advances in the field of information retrieval with verbose queries. Thus, it is clearly relevant to the IR community. The theme of the proposed tutorial is most related to the "Queries and Query Analysis" area of SIGIR 2015. It is also related to a successful series of workshops on "Query Representation and Understanding" held at SIGIR 2010 and 2011.

To the best of our knowledge, no previous tutorials have been offered on this research topic.

5. TOPICS OUTLINE

Here is a brief outline of the topics covered by the tutorial with relevant references.

- Properties of Verbose Queries: Length distribution of queries [2, 45], query types, distribution of mean reciprocal rank wrt query length, part-of-speech distribution, information need specificity [35], repetition factor [37], percentage of searches covered by top-K% queries [37], smoothing mechanisms [46].
- Query Reduction to a Single Sub-Query
  - Types of Sub-Query Candidates: Noun phrases [4], named entities [21], individual words [33, 34], two-term combinations [26], all word subsets [16, 21, 22, 23], word subsets with one word deleted [3, 20], matching queries from personal query log, POS blocks, one to three word queries without stopwords [29], right part of the query [19].
  - Feature Sets
    - Post-Retrieval Features: LambdaRank and BM25 scores of top-K documents [3], query quality, query scope [23], query clarity, standard deviation at 100 documents, a normalised version of standard deviation at 100 documents, the maximum standard deviation in the ranked-list, standard deviation using a variable cut-off point, query length normalized standard deviation using a variable cut-off point.
    - Word Dependency Features: Mutual information between words [21, 22], word co-occurrence in pseudo-relevant documents [29], binary dependencies [33], quasi-synchronous dependencies [34].
    - Query Log based Features: Query log frequency [4], similarity with old queries [20], deletion history, rareness.
    - Statistical Features: TF, IDF [4, 19, 23, 34], simplified clarity score [23], term-term co-occurence, term-topic co-occurence, term-term context, term-topic context [26].
    - Linguistic Features: POS [26, 34], named entities [21, 26], acronyms [26], isBrandName.
    - Query Features: Query length [3, 34], presence of stop words [3, 34], presence of URL [3], similarity with original query [23], isRightMost.
  - Ask for User Input: [21, 22].
  - Methods to Combine the Signals: Word graph construction [21, 22], clustering and rules based approach, classifier [4, 23], LambdaRank [3], stop structure identification [19], rules based approach [20], regression [26], random walk [29], RankSVM [33], quasi-synchronous dependency language model [34].
  - Efficiency Aspect: One word deletion [3, 20], randomly pick up a few sub-queries, overlapping search results and snippets [22].
- Query Reduction by Choosing Multiple Sub-Queries
  - Subset Distributions using CRF-perf [40, 42].
  - Reformulation Trees with Subset Selection and Query Substitution Operations [41].
- Weighting Query Concepts
  - Regression [25, 47].
  - Sequential Dependence Model using Markov Random Fields [30].
  - Integrating Regression Rank with Markov Random Fields (MRFs) [24].
  - Weighted Sequential Dependence Model [7].

  - A Fixed-Point Method [31].
  - Parameterized Query Expansion Model [8].
  - Query Hypergraphs [5].
  - Multiple Source Formulation [9].

- Query Expansion by Including Related Concepts
  - Adding ODP category label to queries.
  - Latent Concept Expansion using Pseudo-Relevance Feedback [8].
  - Selective Interactive Query Expansion [22].
  - Supervised models for pseudo-relevance term selection [12].

- Query Reformulation
  - Translation-based Language Model [43].
  - Random Walks: Query Log-based Term-Query Graph [11], Click Graph [36].
  - Using Anchor Text [15].
  - Question Reformulation Patterns from Query Logs [44].
  - Unified framework for query refinement [18].

- Query Segmentation
  - Statistical Segmentation using Category-wise N-Gram Frequencies [32].
  - Performing Segmentation jointly with Parts-of-speech Tagging and Capitalization [6].
  - Supervised Segmentation using Decision-boundary, Context and Dependency Features [10].
  - Unsupervised Segmentation using Generative Language Models and Wikipedia [38].

- Query-Dependent Learning-to-Rank
  - Query-Dependent Ranking Models [17].
  - Two-Stage Learning-to-Rank Models [14].
  - Learning from Click Data [39, 13].

- Applications of Verbose Query Processing: Finding images for books [1], question answering [19, 43], searching for cancer information, patent searches, fact verification [27], natural language interface for databases [28], e-commerce [32], search queries from children, music search, queries from user selected text.

- Summary and Future Research Directions

Following is a list of related topics that we do *not* cover as part of this tutorial: (1) Automatic Speech Recognition (ASR), (2) Processing null queries other than verbose queries, (3) Query by Document, (4) Query processing tasks for short queries.

6. FORMAT AND DETAILED SCHEDULE

A tentative schedule based on a full-day tutorial is as follows.

- 9:00 – 10:30 Properties of Verbose Queries / Query Reduction
- 10:30 – 11:00 Coffee Break
- 11:00 – 12:30 Query Expansion, Reformulation and Segmentation Techniques
- 12:30 – 14:00 Lunch Break
- 14:00 – 15:30 Weighting Query Concepts
- 15:30 – 16:00 Coffee Break
- 16:00 – 17:30 Query-Dependent Learning-to-Rank, Applications, Summary, and Future Directions

7. TYPE OF SUPPORT MATERIALS

The attendees can download slides from `http://goo.gl/64wY0w` [Draft Version]

No additional equipment is needed.

8. RELEVANT REFERENCES

The following is a list of references which will be used in the preparation of the tutorial material. Many other papers will also be referred.

9. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Enriching Textbooks with Images. In *Proc. of the* 20^{th} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1847–1856, 2011.

[2] A. Arampatzis and J. Kamps. A Study of Query Length. In *Proc. of the* 31^{st} *Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 811–812, 2008.

[3] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring Reductions for Long Web Queries. In *Proc. of the* 33^{rd} *Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 571–578, 2010.

[4] M. Bendersky and W. B. Croft. Discovering Key Concepts in Verbose Queries. In *Proc. of the* 31^{st} *Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 491–498, 2008.

[5] M. Bendersky and W. B. Croft. Modeling Higher-Order Term Dependencies in Information Retrieval using Query Hypergraphs. In *Proc. of the* 35^{th} *Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 941–950, 2012.

[6] M. Bendersky, W. B. Croft, and D. A. Smith. Joint Annotation of Search Queries. In *Proc. of the* 49^{th} *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, pages 102–111, 2011.

[7] M. Bendersky, D. Metzler, and W. B. Croft. Learning Concept Importance using a Weighted Dependence Model. In *Proc. of the* 3^{rd} *ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 31–40, 2010.

[8] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized Concept Weighting in Verbose Queries. In *Proc. of the* 34^{th} *Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 605–614, 2011.

[9] M. Bendersky, D. Metzler, and W. B. Croft. Effective Query Formulation with Multiple Information Sources. In *Proc. of the* 5^{th} *ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 443–452, 2012.

[10] S. Bergsma and Q. I. Wang. Learning Noun Phrase Query Segmentation. In *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, volume 7, pages 819–826, 2007.

[11] F. Bonchi, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. Recommendations for the Long Tail by Term-Query Graph. In *Proc. of the* 20^{th} *Intl. Conf. Companion on World Wide Web (WWW)*, pages 15–16, 2011.

[12] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. In *Proc. of the 31^{st} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 243–250, 2008.

[13] V. Dang, M. Bendersky, and W. B. Croft. Learning to Rank Query Reformulations. In *Proc. of the 33^{rd} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 807–808, 2010.

[14] V. Dang, M. Bendersky, and W. B. Croft. Two-Stage Learning to Rank for Information Retrieval. In *Proc. of the 35^{th} European Conf. on IR Research on Advances in Information Retrieval (ECIR)*, pages 423–434. 2013.

[15] V. Dang and B. W. Croft. Query Reformulation using Anchor Text. In *Proc. of the 3^{rd} ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 41–50, 2010.

[16] S. Datta and V. Varma. Tossing Coins to Trim Long Queries. In *Proc. of the 34^{th} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1255–1256, 2011.

[17] X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum. Query Dependent Ranking using K-Nearest Neighbor. In *Proc. of the 31^{st} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 115–122, 2008.

[18] J. Guo, G. Xu, H. Li, and X. Cheng. A Unified and Discriminative Model for Query Refinement. In *Proc. of the 31^{st} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 379–386, 2008.

[19] S. Huston and W. B. Croft. Evaluating Verbose Query Processing Techniques. In *Proc. of the 33^{rd} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 291–298, 2010.

[20] R. Jones and D. C. Fain. Query Word Deletion Prediction. In *Proc. of the 26^{th} Annual Intl. ACM SIGIR Conf. on Research and Development in Informaion Retrieval (SIGIR)*, pages 435–436, 2003.

[21] G. Kumaran and J. Allan. A Case For Shorter Queries, and Helping Users Create Them. In *Proc. of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 220–227, 2007.

[22] G. Kumaran and J. Allan. Effective and Efficient User Interaction for Long Queries. In *Proc. of the 31^{st} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 11–18, 2008.

[23] G. Kumaran and V. R. Carvalho. Reducing Long Queries using Query Quality Predictors. In *Proc. of the 32^{nd} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 564–571, 2009.

[24] M. Lease. An Improved Markov Random Field Model for Supporting Verbose Queries. In *Proc. of the 32^{nd} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 476–483, 2009.

[25] M. Lease, J. Allan, and W. B. Croft. Regression Rank: Learning to Meet the Opportunity of Descriptive Queries. In *Proc. of the 31^{th} European Conf. on IR Research on Advances in Information Retrieval (ECIR)*, pages 90–101, 2009.

[26] C.-J. Lee, R.-C. Chen, S.-H. Kao, and P.-J. Cheng. A Term Dependency-based Approach for Query Terms Ranking. In *Proc. of the 18^{th} ACM Conf. on Information and Knowledge Management (CIKM)*, pages 1267–1276, 2009.

[27] C. W. Leong and S. Cucerzan. Supporting Factual Statements with Evidence from the Web. In *Proc. of the 21^{st} ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1153–1162, 2012.

[28] Y. Li, H. Yang, and H. Jagadish. Constructing a Generic Natural Language Interface for an XML Database. In *Proc. of the 2006 Conf. on Advances in Database Technology (EDBT)*, pages 737–754, 2006.

[29] K. T. Maxwell and W. B. Croft. Compact Query Term Selection using Topically Related Text. In *Proc. of the 36^{th} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 583–592, 2013.

[30] D. Metzler and W. B. Croft. A Markov Random Field Model for Term Dependencies. In *Proc. of the 28^{th} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 472–479, 2005.

[31] J. H. Paik and D. W. Oard. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In *Proc. of the 23^{rd} ACM Conf. on Information and Knowledge Management (CIKM)*, pages 131–140, 2014.

[32] N. Parikh, P. Sriram, and M. Al Hasan. On Segmentation of E-Commerce Queries. In *Proc. of the 22^{nd} ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1137–1146, 2013.

[33] J. H. Park and W. B. Croft. Query Term Ranking based on Dependency Parsing of Verbose Queries. In *Proc. of the 33^{rd} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 829–830, 2010.

[34] J. H. Park, W. B. Croft, and D. A. Smith. A Quasi-Synchronous Dependence Model for Information Retrieval. In *Proc. of the 20^{th} ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 17–26, 2011.

[35] N. Phan, P. Bailey, and R. Wilkinson. Understanding the Relationship of Information Need Specificity to Search Query Length. In *Proc. of the 30^{th} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 709–710, 2007.

[36] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. LambdaMerge: Merging the Results of Query Reformulations. In *Proc. of the 4^{th} ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 795–804, 2011.

[37] G. Singh, N. Parikh, and N. Sundaresan. Rewriting Null E-Commerce Queries to Recommend Products. In *Proc. of the 21^{st} Intl. Conf. Companion on World Wide Web (WWW)*, pages 73–82, 2012.

[38] B. Tan and F. Peng. Unsupervised Query Segmentation using Generative Language Models and Wikipedia. In *Proc. of the 17^{th} Intl. Conf. on World Wide Web (WWW)*, pages 347–356, 2008.

[39] W. Wu, H. Li, and J. Xu. Learning Query and Document Similarities from Click-through Bipartite Graph with Metadata. In *Proc. of the 6^{th} ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 687–696, 2013.

[40] X. Xue and W. B. Croft. Modeling Subset Distributions for Verbose Queries. In *Proc. of the 34^{th} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1133–1134, 2011.

[41] X. Xue and W. B. Croft. Generating Reformulation Trees for Complex Queries. In *Proc. of the 35^{th} Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 525–534, 2012.

[42] X. Xue, S. Huston, and W. B. Croft. Improving Verbose Queries using Subset Distribution. In *Proc. of the 19^{th} ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1059–1068, 2010.

[43] X. Xue, J. Jeon, and W. B. Croft. Retrieval Models for Question and Answer Archives. In *Proc. of the 31^{st} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 475–482, 2008.

[44] X. Xue, Y. Tao, D. Jiang, and H. Li. Automatically Mining Question Reformulation Patterns from Search Log Data. In *Proc. of the 50^{th} Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 187–192, 2012.

[45] J. Yi and F. Maghoul. Mobile Search Pattern Evolution: The Trend and the Impact of Voice Queries. In *Proc. of the 20^{th} Intl. Conf. Companion on World Wide Web (WWW)*, pages 165–166, 2011.

[46] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models applied to Ad hoc Information Retrieval. In *Proc. of the 24^{th} Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 334–342, 2001.

[47] L. Zhao and J. Callan. Term Necessity Prediction. In *Proc. of the 19^{th} ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 259–268, 2010.