Fusion in Information Retrieval
SIGIR 2018 Half-Day Tutorial

Oren Kurland
Technion — Israel Institute of Technology
Haifa, Israel
kurland@ie.technion.ac.il

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

ABSTRACT
Fusion is an important and central concept in Information Retrieval. The goal of fusion methods is to merge different sources of information so as to address a retrieval task. For example, in the adhoc retrieval setting, fusion methods have been applied to merge multiple document lists retrieved for a query. The lists could be retrieved using different query representations, document representations, ranking functions and corpora. The goal of this half day, intermediate-level, tutorial is to provide a methodological view of the theoretical foundations of fusion approaches, the numerous fusion methods that have been devised and a variety of applications for which fusion techniques have been applied.

1 MOTIVATION
Fusion is a classic technique used for more than twenty years in Information Retrieval, specifically adhoc (query-based) retrieval, that allows multiple sources of information to be combined into a single result set [32, 40]. Fusion can be collection-based, system-based (multiple ranking algorithms), content-based, and even query-based when many similar queries express the same information need [32].

The real power of fusion comes from the fact that even simple aggregation functions have the potential to provide enhanced retrieval effectiveness by exploiting the chorus effect [96].

In this tutorial, we will show that advances in fusion are directly applicable to current open problems in the Information Retrieval community, and that much can be learned from these models as machine learning becomes even more prominent in modern search solutions. In particular we draw parallels between unsupervised fusion and ensembles of classifiers in supervised learning [36, 82, 116].

We focus on retrieval settings where a single corpus is used, and different factors that affect retrieval vary; e.g., queries used to represent the information need, document and/or query representations, ranking functions, etc. We briefly discuss the setting of retrieval over several corpora (a.k.a., federated or distributed search [24, 90]); specifically, we survey several state-of-the-art techniques for fusing lists retrieved from different corpora. We believe that federated search deserves a tutorial in its own right which covers the three main challenges: resource representation, resource selection and results merging [24, 47, 90].

Finally, it is important for everyone in the community to understand just how effective simple fusion techniques can be. Figure 1 and Table 1 compare three state-of-the-art retrieval systems on 100 adhoc queries in the ClueWeb12B UQV100 collection. The three systems being compared are BM25, a field-based SDM model [76] (the exact configuration is identical to the one described by Gallagher et al. [42]), a LambdaMART learning-to-rank (LTR) model [23, 26] (here lightGBM is used with 459 features), and double unsupervised fusion [11, 18] (RRF [29] over all UQV query variations and two systems- SDM-Field and BM25). Figure 1 shows the three strong baselines as a difference in NDCG@10 score w.r.t. a BM25 bag-of-words run. We can clearly see that not only does fusion make more queries better on average, as shown in Table 1, it is also far less likely to make queries worse. This can clearly be seen when comparing Wins, Ties, and Losses (W/T/L) in the table. So, there is much to be learned from fusion baselines when doing exploratory failure analysis on the robustness of new ranking algorithms.

2 TUTORIAL OBJECTIVES
- Highlight the important role of fusion in Information Retrieval.
- Provide a methodological view of the numerous fusion methods.
- Provide an overview of the theoretical foundations of various fusion approaches.
- Introduce the audience to various tasks and challenges for which fusion has been applied and can be applied.
- Discuss parallels with, and more generally pointers to, relevant, related work in machine learning and computational social choice theory.
- Discuss open questions and challenges.

Table 1: Effectiveness comparison of three state-of-the-art ranking methods for the most common query variation for each topic from the ClueWeb12B UQV100 collection [10]. Here ‡ means p < 0.001 in a Bonferroni corrected two-tailed t-test.

<table>
<thead>
<tr>
<th>Method</th>
<th>NDCG@10</th>
<th>W/T/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.212</td>
<td>—/—/—</td>
</tr>
<tr>
<td>SDM-Field</td>
<td>0.233</td>
<td>57/3/40</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.225</td>
<td>59/2/39</td>
</tr>
<tr>
<td>DoubleFuse, v=all</td>
<td>0.300‡</td>
<td>80/1/19</td>
</tr>
</tbody>
</table>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR ’18, Ann Arbor, MI, USA
© 2018 Copyright held by the owner/author(s).
978-1-4503-5657-2/18/07...$15.00
DOI: 10.1145/3289978.3210186
Figure 1: Per topic breakdown comparison of NDCG@10 differences of several state-of-the-art adhoc ranking techniques. The scores shown are the difference between the method and a simple BM25 bag-of-words run. The Double Fusion Technique uses all of the query variations (v=all) for each of the 100 topics, uses RRF Fusion, and combines two systems – SDM-Field and BM25.

3 FORMAT AND PLANNED SCHEDULE

Table 2: Half Day Schedule of Topics

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:15</td>
<td>Introduction</td>
</tr>
<tr>
<td>9:15 - 9:30</td>
<td>Historical Context</td>
</tr>
<tr>
<td>9:30 - 10:00</td>
<td>Theoretical Foundations</td>
</tr>
<tr>
<td>10:00 - 10:30</td>
<td>Fusion in Practice</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Fusion in Practice (contd.)</td>
</tr>
<tr>
<td>11:20 - 11:45</td>
<td>Learning &amp; Fusion</td>
</tr>
<tr>
<td>11:45 - 12:10</td>
<td>Applications</td>
</tr>
<tr>
<td>12:10 - 12:30</td>
<td>Conclusions &amp; Future Directions</td>
</tr>
</tbody>
</table>

OUTLINE

- Intro and Overview
- Historical Context
  - Condorcet, Borda, Kemeny [13, 34, 115]
  - TREC and Rank Fusion [40]
  - Federated Search [24, 90]
- Theoretical Foundations
  - The Fusion Hypothesis [14, 31, 32, 56, 57, 81, 94]
  - Classifier Combination [93]
  - Fusion Frameworks [3, 53, 55, 88, 96, 99, 100, 102]
- Fusion in Practice
  - Score-based (e.g., [3, 40, 56, 57, 78])
  - Rank-based (e.g., [11, 29, 38, 41, 77, 79, 80, 103])
  - Retrieval Score Normalization and Rank-to-Score Transformations [4, 5, 29, 39, 57, 73, 74, 78, 104, 108]
  - Content-based [15, 30, 49–51, 62, 64, 87, 91]
  - Selecting Retrieved Lists for Fusion [43, 45, 46]
  - Query Variations [11, 16–18, 22, 28, 52, 113]
  - Failure Analysis / Risk [18, 37]
  - Efficiency Considerations [44, 59]
- Applications
  - Query Performance Prediction [7, 35, 75, 81, 85, 92, 95, 111]
  - Diversification [60, 63, 109]
  - Relevance Feedback [8, 84]
  - Selecting a Ranker [2, 12, 33, 58]
  - Blog and Microblog Retrieval [60, 61, 64, 101]
  - Pooling and Evaluation [8, 21, 25, 68–71, 86, 97, 98]
- Conclusions & Future Directions

4 TYPE OF SUPPORT MATERIALS TO BE SUPPLIED TO ATTENDEES

- A Web page that contains all materials.
- Downloadable slides available in PDF format.
- Extensive bibliography that helps to further explore topics discussed in the tutorial.
- Scripts and source code for common fusion techniques that can be used by PhD students in future work.

5 PRESENTERS’ BIOGRAPHY

Oren Kurland is an Associate Professor at the Technion — Israel Institute of Technology. He holds a Ph.D. in Computer Science from Cornell University. Oren has served as a senior program committee member and/or area chair for the SIGIR, CIKM, WSDM, WWW and ECIR conferences for the last few years. He is also a member of the editorial board of the Information Retrieval Journal. Oren served as a doctoral consortium co-chair for WSDM 2014, and as a program co-chair for the ICTIR 2013 and SPIRE 2013 conferences. He has also served as the chair of the steering committee of the ACM SIGIR ICTIR conference.

Shane Culpepper is a Vice-Chancellor’s Principal Research Fellow and Associate Professor at RMIT University, and is the Director of the Centre for Information Discovery. His research focuses on building next generation search engines, and exploring new ways to evaluate the quality of search. Research interests include information retrieval, text indexing, data compression, system evaluation, knowledge discovery, machine learning, natural language processing, algorithm engineering and scalability. He is active in many research capacities in the IR research community, including being on the editorial board for the Information Retrieval Journal, and routinely serves on the program committees at ADCS, CIKM, ICDE, SIGIR, SPIRE, WSDM, and WWW. He has been a Program Co-Chair for ADCS, the SIGIR Doctoral Consortium, and CIKM, and will be a General Chair for WSDM 2019.

1https://www.github.com/rmit-ir/polyfuse


