Dynamic Exploratory Search for the Information Retrieval Anthology

Tim Gollub[1], Jason Brockmeyer[2], Benno Stein[2], and Martin Potthast[1]

[1] Leipzig University, first-last@uni-leipzig.de
[2] Bauhaus-Universität Weimar, first-last@uni-weimar.de

Abstract. This paper presents dynamic exploratory search technology for the analysis of scientific corpora. The unique dynamic features of the system allow users to analyze quantitative corpus statistics beyond document counts, and to switch between corpus exploration and corpus filtering. To demonstrate the innovation of our approach, we apply our technology to the IR Anthology, a comprehensive corpus of information retrieval publications. We showcase, among others, how to query for potential PC members and the "Salton number" of an author.

Keywords: IR anthology · exploratory search · faceted search.

1 Introduction and Related Work

The Information Retrieval Anthology[3] compiles a comprehensive collection of publications on information retrieval [6]. At the time of this writing, it includes the bibliographic information of 57,330 IR publications that have appeared since 1963, and indexes the full text of about 88%. It is available online as a search and browsing tool with the goals of (1) providing the information retrieval community with a comprehensive overview of its publications, (2) facilitating scholarly search in a closed-world environment, and (3) enabling community introspection through exploratory and quantitative publication analysis.

While major achievements have already been made and published towards the first two goals, for the first time, this paper reports on our efforts towards the third goal: With "IR Anthology Analytics", we develop a unique exploratory search experience for corpus-based community introspection. Our system design is driven by the assumption that users of the search engine are not merely interested in relevant documents but, beyond that and foremost, in the *analytical statements* that can be made about them.

Existing exploratory search engines, like Relation Browser [1], SearchLens [2], Querium [4], gFacet [5], mSpace Explorer [7], or Flamenco Browser [10], do not embrace this way of thinking, in our view. Though all systems employ, as we do, the concept of faceted search [8] for exploration, facets are meant to serve as document filters much more than they are meant as carrier of analytical statements about the search results.

[3] https://ir.webis.de

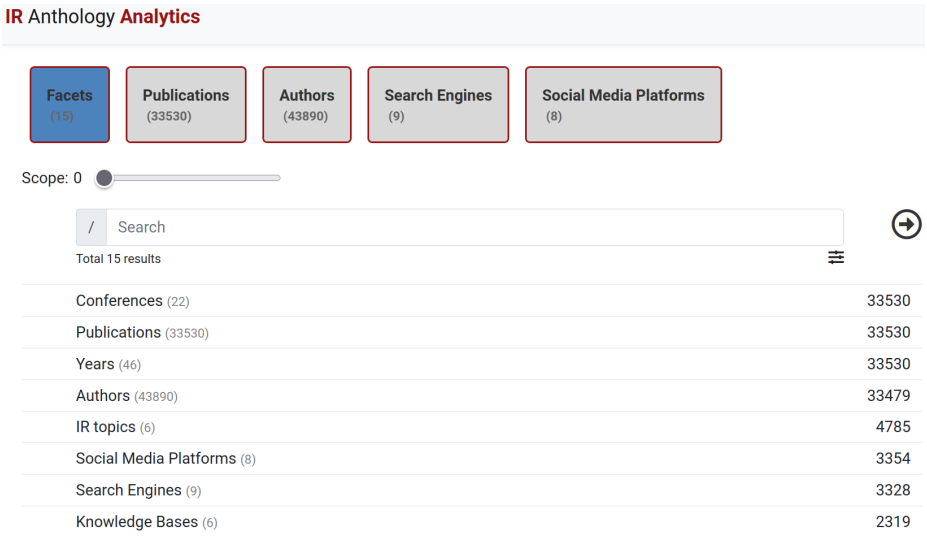

Fig. 1. Screenshot of the presented exploratory search engine. The user interface consists of two main elements. The facet pipe at the top and the facet view below, which shows the terms and term scores of the currently selected facet "Facets" (the root facet, highlighted in blue). The scope of a facet can be set by moving the "scope" slider below the facet pipe.

As a response to this shortcoming, our exploratory search system implements the following three dynamic features: (1) Any facet can be selected as the center of the search results page and hence becomes the target of the search. (2) Relation scores between arbitrary facets can be requested. (3) Both filtering and exploration of the current search results are supported.

The features are described in detail in the following sections. Our current prototype is available at `https://ir-analytics.web.webis.de`.

2 Dynamic Target Facet Feature

As mentioned above, to facilitate a convenient browsing through the terms of any facet, we do not reserve the center of the search results page for the display of the relevant documents. Rather, any facet can be selected to be the current target facet, which is then prominently shown in the center. A screenshot of our user interface, where the root facet "Facets" is the current target facet, is shown in Figure 1. To enable target facet selection, we divide the search results page of our exploratory search engine into two elements, (1) the facet pipe, and (2) the facet view. In the facet pipe, which is displayed at the top of the page, users can add facets relevant to their investigations. Selecting a facet in the pipe shows respective facet terms in a facet view, which is displayed below the facet pipe at the center of the page. For the root facet, the facet view shows all available facets (see Figure 1).

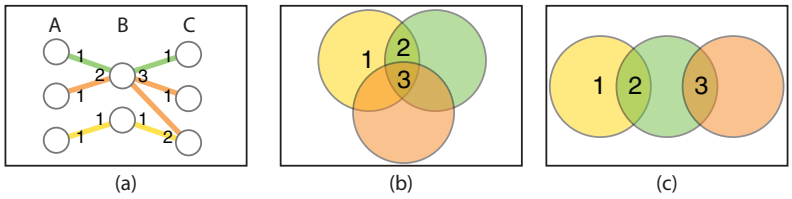

Fig. 2. (a) The numeric value indicators computed for each facet term (= circles) depend on its neighboring facets. E.g., for the upper term of facet B, the left score will be 2 since two terms of facet A are related. The right score will be 3 because of the three relations to terms of facet C. (b) Filter mode. By adding facet terms to the search query, the user can progressively narrow down into smaller result sets (from set 1 to set 3). (c) Exploration mode. By replacing the current query with the selected facet terms, the user can move within overlapping result sets (from set 1 to set 3).

3 Dynamic Scoring Feature

As an innovation to support the statistical analysis of facet term relations, the scores we compute for each facet term of the target facet (called numeric volume indicators in [9]) are not always document counts but depend on the position of the target facet in the facet pipe. The principle is illustrated in Figure 2a. If Facet B is the current target facet, then each of its facet terms displayed in the facet view will feature on the left, the number of related facet terms from Facet A, and likewise, on the right, the number of related faceted terms from Facet C. By moving the position of a facet in the facet pipe, the computation of any term-relationships can be requested. As demonstrated in Section 5, this way, even with a small number of bibliographic facets, interesting statements about the IR community can be made.

4 Dynamic Facet Scope Feature

As pointed out by Gollub et al. in [3], the selection of a facet term by the user can be handled in two different ways: (1) by adding the selected term to the current search query, or (2) by replacing the current search query with the selected term. In the first case, the selected query term is used to filter current search results as illustrated in Figure 2b. In the second case, the selected query term is explored, since the search results feature the relations that this term has to the other facets (see Figure 2c). As both methods have their use cases (see Section 5), by implementing the idea of facet scopes (visible as "scope" slider in Figure 1), the user can decide which method is applied to any facet of the facet pipe.

5 Selected Insights into the IR Anthology

This section illustrates the presented features of our exploratory search engine by linking and discussing the search engine result pages obtained for three selected queries.

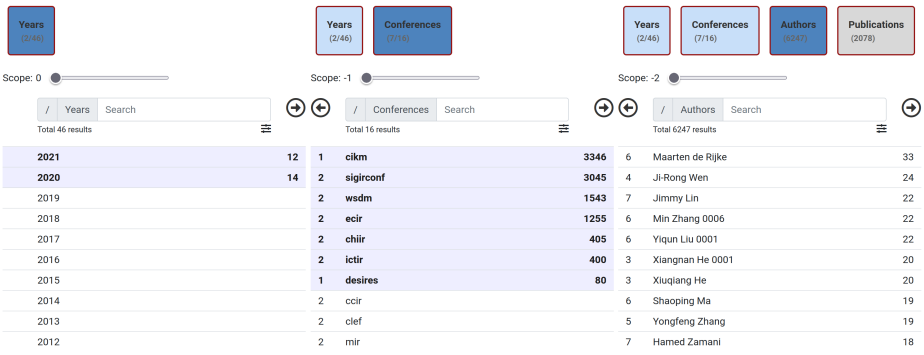

Fig. 3. Three screenshots showing how to construct a search query that reveals the authors (right) who published at any of seven selected IR conferences (middle) in the last two years (left), ordered by number of publications.

5.1 Scoring Module

A common task of conference organizers is to compile a list of active IR researchers which could serve as part of the program committee. To this end, in Figure 3, a collage of three screenshots demonstrating how to query for the authors who published at a major IR conference in the last two years is shown[4]. First, the desired publication years and conferences have been selected in the first two facets. Adding then the authors facet reveals the list of matching authors. By adding a final publication facet, this authors list can be ranked either with respect to number of conferences (left score) or publications (right score).

5.2 Exploration with Scoped Facets

To demonstrate the difference between filtering and exploration, Figure 4 shows a collage of screenshots from the co-author graph of Gerard Salton[5]. Taking inspiration from the Erdős number, which describes the collaborative distance between mathematician Paul Erdős and other persons, the distance of an author to IR pioneer Gerard Salton can be obtained by first adding an authors facet and selecting Gerard Salton, and to then add further authors facets with a reduced scope of -1 (=exploration) to the pipe until the desired author appears in the result. Due to the reduced scope, each authors facet reveals the co-authors of the authors in the previous facet. Note that from the 43 890 authors in the IR Anthology, 34 390 have a Salton number (<=12).

5.3 Content Facets

In order to support explorations of the IR Anthology also with respect to custom facets, we implemented the integration of custom content facets via full-text re-

[4] `https://ir-analytics.web.webis.de/pipes/scoringmodule`
[5] `https://ir-analytics.web.webis.de/pipes/saltonnumber`

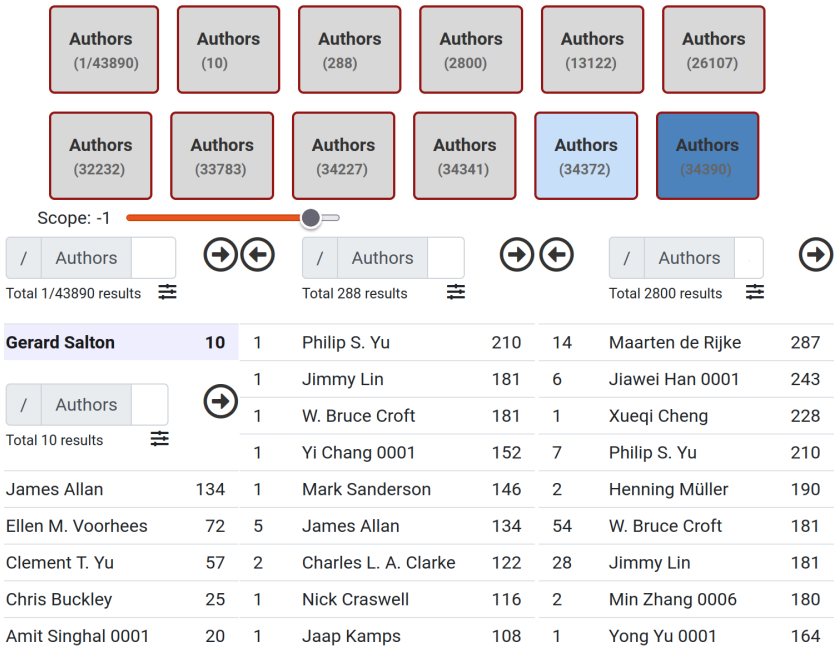

Fig. 4. Collage of screenshots illustrating the co- author graph of Gerard Salton which can be obtained by a series of Authors facets with a scope of -1. The facet views show the result of the first four Authors facets (from top left to bottom right).

trieval. To demonstrate the potential of content facets, we provide a query showing the facet term distribution over publications for a selection of content facets, which we compiled by searching the Web for lists of IR related concepts[6]. The results reveal that Yahoo, Twitter, Wikipedia, and TREC are the most referenced entities in their respective facets Search Engines, Social Media Platforms, Knowledge Bases and Evaluation Forums.

6 Conclusion

This paper reports on our current prototype of an exploratory search engine for the IR Anthology, which we aim to publicly provide to the IR community for introspection. The prototype excels by providing a unique faceted search experience, which is designed especially for investigation tasks. As next steps, among others, we will further extend our scoring module to support relative term counts and non-binary relevance scores. Moreover, we are in the process of crowd-sourcing annotations in the available full-text which pertain to syntactical as well as semantic features. We invite the reader to explore the IR anthology with our current prototype available at `https://ir-analytics.web.webis.de`.

[6] `https://ir-analytics.web.webis.de/pipes/customfacets`

References

1. Capra, R.G., Marchionini, G.: The relation browser tool for faceted exploratory search. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 420–420. JCDL '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1378889.1378967, http://doi.acm.org/10.1145/1378889.1378967
2. Chang, J.C., Hahn, N., Perer, A., Kittur, A.: Searchlens: composing and capturing complex user interests for exploratory search. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019. pp. 498–509 (2019). https://doi.org/10.1145/3301275.3302321, https://doi.org/10.1145/3301275.3302321
3. Gollub, T., Hutans, L., Jami, T.A., Stein, B.: Exploratory Search Pipes with Scoped Facets. In: 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2019). ACM (Oct 2019). https://doi.org/10.1145/3341981.3344247, http://doi.acm.org/10.1145/3341981.3344247
4. Golovchinsky, G., Diriye, A., Dunnigan, T.: The future is in the past: Designing for exploratory search. In: Proceedings of the 4th Information Interaction in Context Symposium. pp. 52–61. IIIX '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2362724.2362738, http://doi.acm.org/10.1145/2362724.2362738
5. Heim, P., Ertl, T., Ziegler, J.: Facet graphs: Complex semantic querying made easy. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications. pp. 288–302. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
6. Potthast, M., Günther, S., Bevendorff, J., Bittner, J.P., Bondarenko, A., Fröbe, M., Kahmann, C., Niekler, A., Völske, M., Stein, B., Hagen, M.: The Information Retrieval Anthology. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021). pp. 2550–2555. ACM (Jul 2021). https://doi.org/10.1145/3404835.3462798, https://dl.acm.org/doi/10.1145/3404835.3462798
7. schraefel, m., Wilson, M., Russell, A., Smith, D.A.: mspace: Improving information access to multimedia domains with multimodal exploratory search. Commun. ACM **49**(4), 47–49 (Apr 2006). https://doi.org/10.1145/1121949.1121980, http://doi.acm.org/10.1145/1121949.1121980
8. Tunkelang, D.: Faceted Search, Synthesis Lectures on Information Concepts, Retrieval, and Services, vol. 1. Morgan & Claypool Publishers (2009). https://doi.org/http://dx.doi.org/10.2200/S00190ED1V01Y200904ICR005
9. Wilson, M.L., m.c. schraefel: mspace: What do numbers and totals mean in a flexible semantic browser (2006), https://eprints.soton.ac.uk/262666/
10. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 401–408. CHI '03, ACM, New York, NY, USA (2003). https://doi.org/10.1145/642611.642681, http://doi.acm.org/10.1145/642611.642681