Time-Based Exploratory Search in Scientific Literature

Silviu Homoceanu, Sascha Tönnies, Philipp Wille and Wolf-Tilo Balke

IFIS TU Braunschweig, Mühlenpfordstraße 23, 38106 Braunschweig, Germany
{silviu, toennies, wille, balke}@ifis.cs.tu-bs.de

Abstract. State-of-the-art faceted search graphical user interfaces for digital libraries provide a wide range of filters perfectly suitable for narrowing down results for well-defined user needs. However, they fail to deliver summarized overview information for users that need to familiarize themselves with a new scientific topic. In fact, exploratory search remains one of the major problems for scientific literature search in digital libraries. Exploiting a user study about how computer scientists actually approach new subject areas we developed ESSENCE, a system for empowering exploratory search in scientific literature.

Keywords: Digital Libraries, User Interface, Exploratory Search, Timeline

1 Introduction

A short survey we conducted among fellow researchers in computer science pointed to a surprising insight: entry points for today's literature search are no longer (digital) library portals, but search engines like Google Scholar or Microsoft Academic Search. Indeed indexing a wide variety of digital libraries, such systems are perfect for exact match searches like looking for a paper, where the title is known, or all recent publications of some author. However, an important part of literature search involves familiarizing oneself with some topic of interest. This kind of search, known as exploratory search, is difficult with any search engine: Starting from rather general keywords one manually has to explore many of the resulting documents and iteratively refine the search in order to get a sufficient overview of the topic. Therefore, in most practical digital library interfaces exploratory search on scientific material is supported through faceted interfaces ([1]). But besides a selection of relevant venues, prolific authors or in the best case, frequent co-occurring keywords for a query, no overview of the field is actually conveyed. Ideally, for a general query like "database systems" exploratory search systems should provide an overview like: In the '60s research focused on hierarchical DBMS, in the '70s relational DBMS were the dominant topic, in the '80s, expert systems and object oriented databases emerged and so on.

We conducted a more detailed user study to analyze how computer scientists approach new subject areas. The main result showed that participants paid special attention to authors' keywords and how they changed over time. Thus, based on sophisticated measures for novelty detection ([2]) we developed ESSENCE (Exploratory Search for SciENCE) a system that extracts emerging keywords from topically focused document collections and presents them on a timeline.

TPDL, p. 1, 2013.

2 User Study - Exploratory Search in Scientific Literature

We conducted a study to understand how scientists become acquainted with a new subject area. All participants (15) had a background in computer science with different levels of expertise, ranging from master students to senior researchers. They were asked to describe their approach on performing literature search on a subject they had low expertise in. All participants proposed to perform a keyword search of this exact term. The tools they used show some differences: While students and young researchers proposed starting with either a Google or Wikipedia search, more experienced researchers chose Google Scholar, Microsoft Academic Research, or Mendeley. The next step they took was to look through the metadata for the found papers: Keywords were generally the first stop, followed by title, time of publication and abstract. Independent of the tools two different strategies were adopted by the participants when exploring an unknown field: The first one was to find "overview" papers. Soon enough it became clear that just one overview paper would not suffice: While early overview papers miss out on what we today refer to as state-of-the-art, recent overview papers focus on the state-of-the-art without covering history or evolution of a field. In consequence, for a complete coverage one needs to consider multiple overview papers published in different time periods. Criteria for identifying overview papers were hints in the paper title e.g., "state-of-the-art", "survey", "overview". However, this approach generates many 'false alarms' while often missing out on actual state-of-the-art papers. The second approach focused on identifying hot topics and how they changed over time. For example, for early papers on "semantic web", keywords like "RDF" or "Metadata" were common. For recently published papers "Linked Data" emerged. Both strategies take important metadata like the keywords and publication time into consideration. But while in the first case participants still had to read at least some of the found overview papers, the second approach already provides an overview by grasping the evolution of keywords over time.

Examining the distribution of keywords' frequencies over time, one can differentiate keywords with high variance in their distribution vs. keywords showing a "flat" distribution. On manual inspection on the results for multiple queries we observed that keywords with low variance in their relative frequency are either general keywords, 'popular' for most fields, or the field itself. In contrast, keywords that were picked up as hot topics for some time interval by study participants, show higher degree of variance over time. They deviate from the expected for the respective time periods appearing more frequent than average. Consistent with the theory of *novelty detection* presented in [2], this observation allowed us to isolate hot topics.

3 System Description

Starting from a paper collection, ESSENCE identifies those papers that are relevant for a scientific field provided as query and extracts those keywords showing high novelty on a yearly basis. Together with other scientific literature metadata they are integrated in summarized form in a GUI (presented in [3]) that facilitates exploratory search. The UI is focused on the two central elements whose importance we identified during the user study: A timeline with query-relevant year span and a tag cloud comprising selected authors' keywords.

A query is a term that best represents a field of interest. For our running example, a possible query would be "semantic web". Throughout this paper we repeatedly make use of the *term* and *document* concepts. By *term* we understand a word or group of ordered words. In accordance with the document metadata that study participants found particularly useful, a *document* is a 5-tuple comprising a document *title*, a set of *keywords* (each keyword is a *term*), an *abstract* a *publication year* and a list of *authors*. Starting with the query term, the system finds those documents that are relevant for the query. A document is a *hit* for a query or any term for that matter, if the term is included in any of the document components: Given a term *t* and a document *d* we define *hit* as a function, *hit* : (Terms × Documents) → {0, 1} with:

$$hit(t,d) = \begin{cases} 1 \text{ iff } t \text{ is contained in } title, keywords \text{ or } abstract \text{ of } d \\ 0 \ otherwise \end{cases} \tag{1}$$

All authors' keywords that annotate documents representing hits for the query are possible feature candidates for the overview. Given a query term q, and a set of documents *H* representing hits for the query term, with $H = \{d \mid hit(q,d) = 1\}$, we define the set of feature candidates for *q* denoted by FC_q as:

$$FC_q = \{k \mid k \ is \ a \ keyword \ of \ d, \forall \ d \in H\}. \tag{2}$$

Publication dates are also extracted in the process since they are needed for computing the publication time span (year of publication of the earliest published paper - year of publication of the latest published paper) for the query. Given a query *q*, we define the *query years set* for *q* denoted by Y_q as:

$$Y_q = \{y \mid |D_{q,y}| \geq \theta\}, \tag{3}$$

where $D_{q,y} = \{d \mid hit(q,d) = 1 \land d \ was \ published \ in \ year \ y$. The lower the value of θ the less significant are the resulting estimations.

For each year in the publication time-span, a yearly term weight is computed for all extracted keywords: Given a query q, a publication year *y*, and a term *t* we define the *yearly term weight* for term *t* under query *q* in year *y*, denoted by $w_{q,y}(t)$ *as* a function, $w_{q,y}(t)$: Terms → [0, 1] with:

$$w_{q,y}(t) = \begin{cases} \frac{1}{|D_{q,y}|} \cdot \sum_{i=1}^{|D_{q,y}|} hit(t,d_i), with \ d_i \in D_{q,y}, \ iff \ D_{q,y} \neq \emptyset; \\ 0, \qquad\qquad otherwise; \end{cases} \tag{4}$$

where $D_{q,y} = \{d \mid hit(q,d) = 1 \land d \ was \ published \ in \ year \ y\}$. (5)

The yearly term weight seems like a good mechanism for determining which feature candidates show high weight variance. But computing a measure of variance like the standard deviation of the weights for each feature candidate favors mainstream keywords: The standard deviation of mainstream keywords is much bigger than the standard deviation of more specific, lower frequency keywords. Despite representing important features, keywords with low frequencies would never be considered relevant. For this purpose, normalizing the standard deviation by the average (known as the coefficient of variation) is necessary: Given a query *q*, and a set of feature candidates FC_q for *q*, we define the set of features for *q* denoted by F_q as:

$$F_q = \{f \mid f \in FC_q \wedge \frac{stdev\left(\bigcup_{i=1}^{|Y_q|} w_{q,y_i}(f)\right)}{avg\left(\bigcup_{i=1}^{|Y_q|} w_{q,y_i}(f)\right)} \geq \gamma\}, with\ y_i \in Y_q, \tag{6}$$

where *stdev* and *avg* represent the standard deviation and average of all weights for *f* under query *q* and γ regulates the lowest acceptable frequency distribution.

F_q comprises a list of features that are relevant for query *q*, but the relevance of each feature on a yearly basis still has to be determined. The *yearly term weight* is not suitable since it favors mainstream features. Instead, a function that captures the normalized positive deviations on a yearly basis is necessary: Given a query *q*, a publication year *y*, and a term *t*, we define the *yearly term novelty* for term *t* under query *q* in year *y*, denoted by $n_{q,y}(t)$ *as* a function, $n_{q,y}(t)$:Terms → [0; ∞) with:

$$n_{q,y}(t) = \begin{cases} \frac{w_{q,y}(t) - avg_{w_q}(\mathrm{t})}{avg_{w_q}(t)}, & iff\ w_{q,y}(t) > avg_{w_q}(\mathrm{t}) > 0; \\ 0, & otherwise. \end{cases} \tag{7}$$

$$\text{where } avg_{w_q}(\mathrm{t}) = avg\left(\bigcup_{i=1}^{|Y_q|} w_{q,y_i}(t)\right), with\ y_i \in Y_q. \tag{8}$$

Finally, the relevance of features (from F_q) for a given query *q* over time, is computed as the *yearly term novelty* of the feature for the corresponding relevant years (from Y_q).

4 Conclusion

Systems like Google Scholar, or Microsoft Academic Research, favoring simple yet effective interfaces are the first stop when searching for literature in computer science. However, for exploratory search, even the more sophisticated faceted search interfaces don't perform to the users' satisfaction. Learning from the way users familiarize themselves with new scientific areas, we discovered that their approach is consistent with the theory of novelty detection successfully implemented in online news mining. ESSENCE adapts these techniques to the particularities of scientific literature for extracting overview information.

5 References

1. J. Diederich and W.-T. Balke, "FacetedDBLP - Navigational Access for Digital Libraries", *TCDL*, 2008.
2. J. Ma and S. Perkins, "Online Novelty Detection on Temporal Sequences", *KDD*, 2003.
3. S. Homoceanu, S. Tönnies, P. Wille, W.-T. Balke, "ESSENCE- Time-Based Exploratory Search in Scientific Literature", http://dx.doi.org/10.6084/m9.figshare.710918, 2013.