DiLiA – The Digital Library Assistant

Kathrin Eichler, Holmer Hemsen, Günter Neumann, Norbert Reithinger, Sven Schmeier, Kinga Schumacher, and Inessa Seifert

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
firstname.lastname@dfki.de,
DFKI home page: http://www.dfki.de
DiLiA home page: http://dilia.b.dfki.de/

Abstract. In this paper we present the digital library assistant (DiLiA). The system aims at augmenting the search in digital libraries in several dimensions. In the project advanced information visualisation methods are developed for user controlled interactive search. The interaction model has been designed in a way that it is transparent to the user and easy to use. In addition, information extraction (IE) methods have been developed in DiLiA to make the content more easily accessible, this includes the identification and extraction of technical terms (TTs) – single and multi word terms – as well as the extraction of binary relations based on the extracted terms. In DiLiA we follow a hybrid information extraction approach – a combination of metadata and document processing.

1 Introduction

Although the content of digital libraries is growing rapidly, popular portals for digital libraries, such as Google Scholar, Citeulike, ACM digital library still limit the search options to a small set of meta labels (such as author, title, etc.) and only provide a limited text-based search interface. So far, these portals do not use any elaborated visualisation techniques for presenting the search results. This is problematic in two ways. Firstly, since the search options are restricted to metadata, a search query that is not specific enough will easily lead to a long list of search results. Secondly, since no elaborated visualisation techniques are used, navigating through the search result is difficult and time consuming. The goal of DiLiA is to go beyond this level of information access. We especially target users that want to interactively explore the content of the digital library, for example, users that want to investigate a new research area. The DiLiA demonstrator is based on real data in the computer science domain. The database contains 1.2 million abstracts with corresponding metadata from DBLP.

2 Visualisation

The development of the user interface has been led by the design principle that the visual representations should provide clues: what can be done next and what are the possible directions for further search [1]. The user interface consists of a

relational view, visualising relations between the search queries the user specifies; a *hyperlink activated list of search results* with detailed information on each item; and *tools* (e.g., bar charts) for a flexible analysis of the search results.

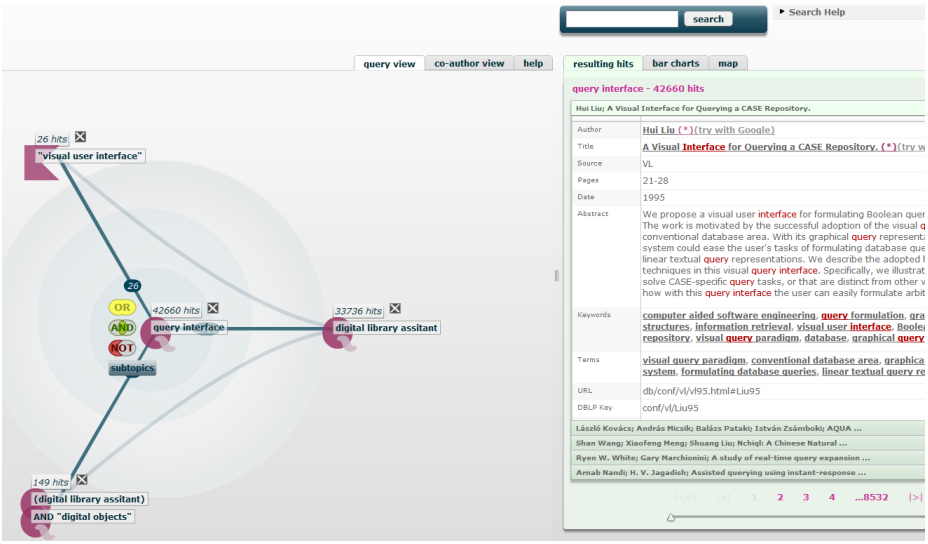

Fig. 1. User interface of the DiLiA demonstrator

Figure 1 shows the user interface (UI) of the DiLiA demonstrator. On the right side of the UI the search panel with the search result list is located. Selecting an item in the result set shows its metadata. Users have different possibilities for stating a search query. The search panel allows the user to enter a search term for searching in the digital libraries metadata. In addition, the user can add items from the hyperlink activated search result list, e.g., a keyword or an author name. On the left side of the UI relations between search queries are displayed (TopicView) in form of a graph (TopicGraph). Each search topic is represented as a TopicBlob (node). Edges between the TopicBlobs show the number documents common in both TopicBlobs. The TopicBlobs can be combined interactively via drag-and-drop on boolean operators (AND, OR, NOT) included in the TopicView (for details see [2]). For each TopicBlob in focus, subtopics can be viewed and selected. The subtopics are automatically generated by dynamically clustering the document abstracts using the Carrot^2 Clustering engine[1]. On the right side of the UI the user can also switch to various views, supporting visual analytics on the data. The bar chart view shows how many documents have been published in a specific year for the selected topic. Depending on the curve that the bar chart forms the user might be able to see if a research topic is a hot topic or if few papers have been published on the topic lately. The user has also

[1] http://project.carrot2.org/

the possibility to use a heatmap view. The heatmap shows using a world map the origin of the publications about the topic and how it emerged over time and enables the user to see where in the world a research topic started and how it spread. On the left side of the UI the user can switch to an author graph, showing for a selected publication the author, the co-authors and the publications of author and co-author and to navigate further.

3 Information Extraction

The goal of IE in DiLiA is on one hand to support digital libraries in the process of making available new material and on the other hand to support users in interactively exploring the content. We have developed a Generalised Name Recogniser (GNR) for identifying domain independent, fully automatically and unsupervised, multi-word technical terms, cf. [3]. Processing only the abstracts of the documents, the current prototype contains these technical terms as automatically generated list of keywords. Based on the identification of TTs in the whole document, we are currently working on unsupervised relation extraction methods. The extracted relations can be used for advanced search and also serve as basis for clustering similar relations/documents. For identifying the TTs[2] we used the nominal group (NG) chunker of the GNR, but the output was modified. For example, coordinated phrases had to be split or text in parenthesis had to be processed separately [3]. Since not every NG is a TT, we needed to find a way to filter the NGs. Inspired by Luhn's findings [4], who suggested that mid-frequency terms are the ones that best indicate the topic of a document, frequency scores for all NGs using the Live Search API from Microsoft are retrieved. The NGs are then filtered using an upper and lower threshold. We found out that the upper threshold is domain dependent. For computer science documents the best F-measure was achieved with a threshold of 20 mio., for biology 6.5 mio. The extracted TTs serve as the basis for relation extraction.

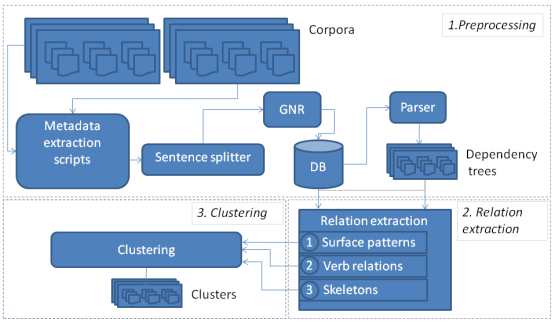

Fig. 2. Information extraction data flow in DiLiA

[2] An evaluation on a hand-annotated computer science corpus (DBLP) showed that 68.2% of the NGs were identified completely and 31.3% partially (caused by missing prepositional postmodifiers, additional premodifiers and appositive constructions).

Fig. 2 shows the information flow. For the IE process all documents are first split into sentences. The identified TTs are then replaced in each sentence with a termID. Three different binary relation strategies have been implemented and are currently being evaluated. The first strategy "surface patterns" is inspired by [5] and uses the following pattern <TermID1>string<TermID2> to match each sentence against. For "Verb relations" and "Skeletons" the modified sentences are parsed with the Stanford Parser with dependency tree output. In the "Verb relation" IE method the verb node and direct neighbour nodes containing TTs are extracted. In the "Skeleton" approach [6] the relation consists of information collected by going up the dependency tree starting from pairs of TTs and ending at a common root node.

4 Conclusion and Future Work

In this paper we presented the DiLiA demonstrator[3], which provides a novel user interface for interactively navigating in a digital library database. The system also integrates IE methods (automatic extraction of technical terms and binary relations). Currently, we are working on the implementation of the DiLiA system for a Touchmaster touch table (2 x 1.10m) and investigate clustering algorithms for very large data sets.

Acknowledgment

The research project DiLiA is co-funded by the European Regional Development Fund (ERDF) in context of Investitionsbank Berlin's ProFIT program under grant number 10140159. We gratefully acknowledge this support.

References

1. Marchionini, G.: Information-seeking strategies of novices using a full-text electronic encyclopedia. J. Am. Soc. Inf. Sci. **40**(1) (1989) 54 – 66
2. Seifert, I., Kruppa, M.: A pool of topics: Interactive relational topic visualization for information discovery. In Huang, M.L., Nguyen, Q.V., Zhang, K., eds.: Visual Information Communication, Springer (2010)
3. Eichler, K., Hemsen, H., Neumann, G.: Unsupervised and domain-independent extraction of technical terms from scientific articles in digital libraries. In Mandl, T., Frommholz, I., eds.: Proc. of the Workshop "Information Retrieval", organized as part of LWA, Darmstadt, Germany (21-23 September 2009)
4. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development **2** (1958) 157–165
5. Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., Ishizuka, M.: Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from theWeb. In: Proc. of ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore (2009) 1021–1029
6. Wang, R., Neumann, G.: Recognizing textual entailment using a subsequence kernel method. In: Proc. of AAAI-2007, Vancoucer, Canada (2007)

[3] Online accessible via: `http://dilia.b.dfki.de/`