

Hajra, Arben; Tochtermann, Klaus

Conference Paper — Accepted Manuscript (Postprint)

Visual Search in Digital Libraries and the Usage of External Terms

Suggested Citation: Hajra, Arben; Tochtermann, Klaus (2018) : Visual Search in Digital Libraries and the Usage of External Terms, In: 2018 22nd International Conference Information Visualisation (IV), Fisciano, Italy, 10-13 July 2018, IEEE, New York, pp. 396-400, <https://doi.org/10.1109/iv.2018.00074>

This Version is available at:

<https://hdl.handle.net/11108/393>

Kontakt/Contact

ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
<https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw>

Standard-Nutzungsbedingungen:

Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

Terms of use:

This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.

Visual Search in Digital Libraries and the Usage of External Terms

Arben Hajra

Leibniz Information Centre for Economics (ZBW)
Kiel/Hamburg, Germany
a.hajra@zbw.eu

Klaus Tochtermann

Leibniz Information Centre for Economics (ZBW)
Kiel/Hamburg, Germany
k.tochtermann@zbw.eu

Abstract— This paper focuses on the application of visual search interfaces in the context of digital libraries. The main objective is to represent a simplified and intuitive interactive approach for retrieving similar publications based on a preselected one. This would enable the scholar to perform more detailed research with the reduced mental workload, in comparison to traditional keyword-based search. The proposed approach, in an innate and conceptual manner, makes possible the application of suggested terms from other external resources. Accordingly, the set of terms can be extended with synonyms, narrowed, broadened or closely related terms. Such suggestions may result from a simple language thesaurus, any SKOS modelling scheme, and the deployment of word embedding approaches, such as word2vec. To provide a better picture of why a particular publication is presented in the results list, the matched terms are colored.

Keywords— tag cloud, visualization, visual search, thesaurus, word embedding

I. INTRODUCTION

Digital libraries (DLs) represent important knowledge in the scientific community. As such, they are a crucial part of the everyday activities of any scholar. However, search interfaces offered by current DLs are not always suited for retrieving what scholars are looking for [1]. Such interfaces principally are based on search terms, which initially are provided by the scholar, often in combination with various other facets for narrowing down the search result. For the sake of better results, the scholar should selectively and carefully reflect about the (combination of) terms. Accordingly, this search strategy may possibly affect the limitation of the search only to the terms set forth at the beginning. Therefore, one of the main aims presented in this work is to introduce an approach that would facilitate the search process based on the terms extracted from an already retrieved publication.

This paper focuses on the application of a visual search interface in the context of DLs, i.e. scientific publications in DLs. By highlighting the metadata of a publication, the scholar can retrieve semantically similar publications close to that publication. The deployment of several visualization elements, such as colors, shapes, word clouds, sliders, etc., increases the accessibility and interactivity of the results provided. Thus, the scholar is provided with a satisfactory result using a slightly simpler route. In addition, every step taken is completely intuitive and cognitive at the same time.

Furthermore, the set of terms extracted from the publication's metadata (title, abstract, keywords) can be enriched with several other terms through the deployment of

external resources. Thus, it is possible to include concepts from any (connected) external language thesauri, or for a comprehensive hierarchical navigation within any SKOS modelling scheme [2]. As a result, the set of already existing terms are extended with synonyms, narrowed, broadened or closely related terms. In addition, the approach enables the adoption of terms generated through machine learning techniques, i.e. word embedding approach (Word2Vec) [3]. Through a well-analyzed approach, a range of functionalities can be achieved without having to overload the design and interactivity. Adding all these concepts gives an ultimate control to the scholar for making the anticipated adjustment for getting the best out of a DL.

II. RELATED WORK

The visual search paradigm is not new. It is actually one of the most gainful examples related to the deployment of attention, where the subject attempts to find the target among the set of distractors, such as shape, color, or size [4]. In addition, the study in [5] tackles the importance of information visualization for visual search, especially in DLs. The use of color, size, shape, orientation, position, organization and relations in content, are important elements for increasing the user attention. However, the deployment of such elements requires an appropriate strategy for interweaving metadata and the visual mode to improve the access to digital resources [5], [6].

Several other studies also put an emphasis on the position of visual search in the context of DLs, i.e. searching text documents. According to [6]–[9], such application facilitates the exploration and navigation processes through different resources. Moreover, the scholar can complete the search with reduced mental effort [8]. Interesting insights regarding the needs, requirements and challenges of visual search in DLs are presented by Kitchenham, et al. [9].

The hierarchical navigation through the Simple Knowledge Organization System Reference - SKOS modelling scheme, inside a DL, is a crucial operation in terms of narrowing the obtained results [2], [10]–[12]. However, browsing the taxonomic levels increases the complexity of the interactivity and usability of any search interface [13]. Therefore, the application of visual search expressively affects navigation simplicity, but also increases the potential for maximum utilization [14]. The advantages of visual search over text-keyword queries, except on text-based collections, are also evident in other domains such as in video content or other digital resources [15].

One of the most used elements for pure text summarization is word cloud, known also as a tag cloud. As noted in [16], [17], word clouds are used in various contexts as a means to provide an overview by visualizing the text to those words that appear with a particular weight. Using word clouds for depicting the representative keywords is also shown in [18]. In most of the cases, the weight represents the frequency of that word in a given text (term frequency - tf). However, other calculations that define the relevance of the terms in a given corpus are also possible. The multiplication of tf with the inverse document frequency idf , known as $tf-idf$, is also a very popular way to calculate the word's weight.

III. MOTIVATION

Let us consider a scenario. We want to find a similar song to what we just heard, but with few different features, such as with lower rhythm, and a more dominant piano. Or we like to get more movies similar to what we have seen today, but with fewer scenes of violence, more dramatic and very mysterious. Both scenarios have in common that they seek for a product which is recomposed of similar products by selecting features that we like or dislike.

In the context of scholarly communication, the need for something like this is obvious. Let us assume that we have found an interesting publication in our favorite DL, entitled "*Globalization, brain drain and development*". This DL also offers a list of suggested publications based on it. However, what if we prefer a list of recommended publications which are more related to "*brain drain*" rather than "*globalization*"?

The above scenarios mainly focus on the user, i.e. scholar behavior for discovering and consuming publications. In most common cases, a scholar refers to DLs to research. Therefore, after some searches she has found a publication that complies with her request, and is interested in publications comparable to it. Principally, almost every DL provides such a list of feeds, i.e. recommending based on a selected publication. For example, Google Scholar¹ offers the option "Related Articles", Mendeley² has "Suggestions Based on This Article", EconBiz³ "Similar Items by Subject" while Elsevier's ScienceDirect⁴ offers "Recommended articles", etc. An in-depth overview for facilitating faceted search is provided by the EEXCESS⁵ project. However, from what we have observed, most of the existing approaches lack the opportunity for a detailed customization of the recommended publications, with the purpose of specifying the results. In addition to common layouts for narrowing down the results, when multiple functionalities are applied, there is an overload of the designs. For example, this becomes evident when a scholar is not aware why a particular publication appears in the result list. This is especially the case when an external thesauri is used as to expand a scholar's search terms. Therefore, she remains unaware about the presence of that term in the query formulation and publications retrieval. Within this context, our approach tends to introduce a balanced interface between simplicity and functionality, i.e. getting more with less effort.

IV. THE APPROACH

Our intention is related to the opportunity of the scholars to be able to re-select the recommended publications by redefining the concepts of the selected publication. To facilitate the process, the representative words (terms, keywords) assigned to a publication are extracted and visualized. Such visualizations provide the scholar with an instant overview of concepts related to a publication.

For each paper selected by the scholar, the system determines the key concepts based on the retrieved metadata of a publication. Therefore, the scholar can adjust metadata and weights of concepts to narrow the results.

Our approach has been developed and assessed with the content of the EconStor repository, a leading Open Access repository in Germany [19]. Through EconStor, the Leibniz Information Centre for Economics (ZBW) offers a platform for Open Access publishing to researchers in economics. The repository metadata is accessible through a portal, a SPARQL Endpoint, or as RDF triples dump file. ZBW also maintains the Standard Thesaurus Wirtschaft (STW), which is the Thesaurus for Economics used for description and indexing purposes [20].

A. Searching

A detailed overview of the approach and the interaction interface is given below. Figure 1 shows a scenario in which a scholar selects a particular publication, for which its metadata are projected in the word cloud. As depicted, the scholar's search activity is concentrated on three main areas, in order to advance her visual search and retrieving other closely related publications. Through the area 1, as denoted in figure 1, the scholar can determine which metadata components to consider. There are three components in total: title, abstract and keywords. Thus, the scholar has the possibility to include or exclude any of them at the same time to determine the importance for each of them, by increasing or changing its value. As seen in the example of figure 1, the title is factorized more than other elements. Based on our previous work [21], as the most determinant combination, we have perceived the combination of all of them by doubling the weight/importance/impact of the title. The title is often most representatives, as authors tend to include the key terms regarding the subject in it.

The output of metadata combinations from area 1 is visible immediately within the tag word cloud in area 2. The application of $tf-idf$ emphasizes the importance of the terms in the word cloud. The main interactivity action is taken through the usage of the slider in the area 2a. Thus, its movement defines the number of terms to be taken into account in the word cloud, i.e. search terms. The combination of all metadata elements (title, abstract, keywords) may produce a large set of terms. Hence, the user can determine the percentage of terms to be considered, starting from the most important. All this influences the generated results since the user can determine the presence or absence of the less important terms at further calculations. As noted, the first interactivity between the scholar and the interface, regarding the retrieved results, is exactly the slider in area 2a. Each interactivity with that slider results in changes considering the ranking of recommended publications.

¹ <https://scholar.google.com/>

² <https://www.mendeley.com/>

³ <https://www.econbiz.de/>

⁴ <https://www.sciencedirect.com/>

⁵ <http://eexcess.eu/visualisations/>

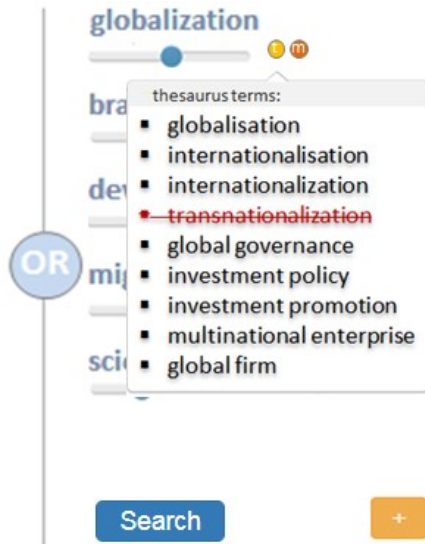


Figure 3. Terms suggested by the thesaurus.

Machine Learning Terms - are generated through the deployment of a word embedding approach. The vector representations of words over neural networks, i.e. word embedding, represent one of the most trending topics linked to word relatedness and similarities [23]. Such word representations are produced as a result of the trained model on a large collection of text corpora. Therefore, in the continuous vector space model words are embedded in relation to their semantic similarity. Among several word embedding techniques, we have applied Word2Vec algorithms proposed by Mikolov et al. for Google [24], [25]. The Python implementation of Word2Vec, based on the Gensim package, is part of our implementation.

At this time, there exist several pre-trained models on different datasets, such as Google News, Wikipedia+Wikipedia+Gigaword, Twitter, DBpedia, and Freebase. For our purpose, we have based our experiments on two models: Google News and the model trained by ourselves.

The Google News model is trained on 100 billion words from the Google News dataset. As such, the model contains term vectors of 3 million terms including phrases. The terms inside the vectors are distributed in a 300-dimensional dimensionality space, which means that each term is represented with 300 most similar words in that vector. On the other side, our own model is trained with the EconStor content, by considering the title and abstract of all publications. Hence, the model is trained on a corpus of around 12 million words, with a windows size of 5 and 300 dimensions. Based on our previous work [21], building a model on top of a specific domain-related datasets, the Word2Vec gives closely related domain correlations of terms; whereas the existing pre-trained models such as Google News provides more general context to a particular term.

Figure 4 shows an example of Word2Vec implementation for suggesting top five most similar terms, considering the term “*globalization*”. The suggestions came from the model that we have created during our research. For the same word, the Google News model categorise the following terms as most similar to “*globalization*”:

globalism, globalized, globalizing, globalization and capitalist globalization (not shown in figure 4).

As explained in the section “Thesaurus Terms”, the scholar furthermore may control the presence of these concepts in the searching process. At any time, she can exclude any of them or the complete list from further computations. In addition, she can extend the list of suggestions with five new words (+5). Hence, referring to our model, the list of suggestions for the word “*globalization*” will be extended with *integration, liberalisation, deep, global and resilience*.

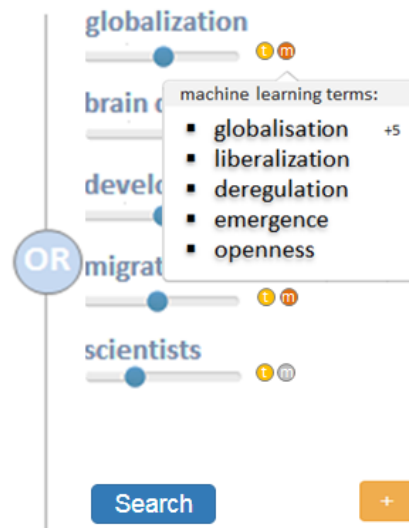


Figure 4. Terms suggested from machine learning techniques.

V. THE SIMILARITY MEASUREMENT

Retrieving and ranking the list of publications, based on the scholar’s search input is done by measuring the angles between vectors of concepts through the Cosine Similarity. In principle, the measurements are made between the terms invoked by the initial publication (p) and the terms of each publication in the respective repository ($d_i, i = 1, n$).

Accordingly, in the initial side, there are all terms that the user has selected from the publication p . There can be all the terms appearing in area 2 of figure 1, or the set of preselected terms from area 3. In the second case (from area 3), the set also includes the terms suggested by the thesaurus and machine learning techniques. On the other side, the vectors d_i contains the terms of publications to be measured for calculating the similarity degree. In total, there are n comparisons, that in fact is the total number of publications in that repository D . If we are performing at the same repository, the initial publication is excluded, ($n=|D|-1$). Thus, iteratively, as described in [8], we measure the similarity between metadata of our initial publication (p) with the metadata of publications from the target repository (D), i.e. $sim(p, d_i)$, for $i=1, n$.

VI. THE OUTCOME

Our results offer the scholar with the possibility to operate with several sets of terms, including the external suggestions from thesauri or machine learning. Hence, in some cases, it may be difficult for the scholar to recognize

Financial liberalization and the brain drain: A panel data analysis

This paper explores the impact of financial liberalization on the migration of ... financial liberalization, namely the robustness of the markets and their freedom from direct control ... the emigration of high skilled labor and the effect is not statistically significant ... economic globalization ...

Subjects: financial liberalization, brain drain, institutions, immigration

Figure 5. Retrieved results.

why a particular publication is displayed on the list of results, based on a particular search.

For better interpretation of the results, particularly, to have a clear picture of why a particular publication is presented in the results list, the matched terms are colourized. Figure 5 gives an overview of such a visualization. The black bolded text represents the match of the terms from area 3; the red text shows that the match came from the machine learning suggestion; and the yellow colour represents terms from the thesaurus suggestions. The mouseover event also displays a popup notification about the source of the matched term.

VII. CONCLUSION

In this paper we represent a simplified approach for applying a visual search interface in the context of digital libraries. Contrary to traditional forms of searching, i.e. keyword-based input queries, the proposed approach attempts to increase the possibilities for achieving better search results by reducing the mental engagement of the scholars by providing a simpler and colored interface. Therefore, from a single search interface, the scholar can perform several functionalities for satisfying her search. The introduced customization of attributes is very intuitive and easily applicable. The application of external thesauri and suggested terms through machine learning techniques are applied innately. This allows the scholar at any time to manage the features and instantly see the change reflected on the results.

The proposed approach is currently being evaluated. So far, we have informally evaluated it with five scholars (master students), to perceive the user experience, interactivity and functionality. Initial impressions and feedbacks of are very positive; however, we expect to learn more once the evaluation is completed.

REFERENCES

- [1] D. Martin-Moncunill, P. A. Salvador Sánchez Alonso, G. García, and N. Marianos, "Applying visualization techniques to develop interfaces for educational repositories," *Ling. VOA3R*, p. 60, 2013.
- [2] A. Isaac and E. Summers, "SKOS Simple Knowledge Organization System," *Prim. World Wide Web Consort.*, 2009.
- [3] Y. Goldberg and O. Levy, "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv Prepr. arXiv1402.3722*, 2014.
- [4] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychon. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.
- [5] P. A. Gaona-García, D. Martin-Moncunill, and C. E. Montenegro-Marin, "Trends and challenges of visual search interfaces in digital libraries and repositories," *Electron. Libr.*, vol. 35, no. 1, pp. 69–98, 2017.
- [6] M. Hearst, *Search user interfaces*. Cambridge University Press, 2009.
- [7] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted visualization for rich text corpora," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1172–1181, 2010.
- [8] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval: The Concepts and Technology behind Search (ACM Press Books)." Addison-Wesley Professional Harlow, 2011.
- [9] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—a systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.
- [10] Z. Wang, A. Sattar Chaudhry, and C. S. G. Khoo, "Using classification schemes and thesauri to build an organizational taxonomy for organizing content and aiding navigation," *J. Doc.*, vol. 64, no. 6, pp. 842–876, 2008.
- [11] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau, "Visualizing digital library search results with categorical and hierarchical axes," in *The Craft of Information Visualization*, Elsevier, 2003, pp. 169–177.
- [12] M. Nasir Uddin and P. Janecek, "Performance and usability testing of multidimensional taxonomy in website search and navigation," *Perform. Meas. metrics*, vol. 8, no. 1, pp. 18–33, 2007.
- [13] G. Marchionini and R. White, "Find what you need, understand what you find," *Int. J. Hum. [x02013] Comput. Interact.*, vol. 23, no. 3, pp. 205–237, 2007.
- [14] P. A. Gaona-Garcí, G. Stoitsis, S. Sánchez-Alonso, and K. Biniari, "An Exploratory Study of User Perception in Visual Search Interfaces Based on SKOS," *Knowl. Organ.*, vol. 43, no. 4, 2016.
- [15] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Reranking methods for visual search," *IEEE Multimed.*, vol. 14, no. 3, 2007.
- [16] M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. Di Silvestro, and D. Weiskopf, "RadCloud: Visualizing multiple texts with merged word clouds," in *Information Visualisation (IV), 2014 18th International Conference on*, 2014, pp. 108–113.
- [17] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 2014, pp. 1833–1842.
- [18] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in *Visualization Symposium (PacificVis), 2010 IEEE Pacific*, 2010, pp. 121–128.
- [19] A. Latif, T. Borst, and K. Tochtermann, "Exposing data from an open access repository for economics as linked data," *D-Lib Mag.*, vol. 20, no. 9/10, 2014.
- [20] J. Neubert, "Bringing the ' Thesaurus for Economics ' on to the Web of Linked Data," vol. 25964, 2009.
- [21] A. Hajra and K. Tochtermann, "Linking science: approaches for linking scientific publications across different LOD repositories," *Int. J. Metadata, Semant. Ontol.*, vol. 12, no. 2/3, pp. 121–141, 2017.
- [22] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [23] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.
- [24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.
- [25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Prepr. arXiv1301.3781*, 2013.