An Investigation into Information Navigation via Diverse Keyword-based Facets

M. Atif Qureshi and Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland
{muhammad.qureshi@ucd.ie, derek.greene@ucd.ie}

Abstract. In the age of information overload, it is necessary to provide effective information navigation tools that operate over unstructured textual data. Current state-of-the-art methods are limited in terms of providing effective exploration capabilities for various information seeking tasks, especially those arising in domains such as online journalism. Here we argue for improvements in faceted search systems, via new strategies for identifying keyword-based facets. Our proposed technique utilises a PageRank model operating over the graph of terms appearing in documents, while also employing novel methods for biasing significant terms and named entities. In addition, we consider the notion of diversity within extracted keywords in an effort to maximize coverage over a range of topics. We perform experimental evaluations over issues relevant to the Irish General Elections 2016, demonstrating the superior performance of our proposed technique.

1 Introduction

Web 2.0 technologies have enabled online information to grow at an exponential rate, with a sizable form of this information being textual in nature. To a large extent, the textual information is unstructured, leading to what is commonly known as "information overload" problem for lengthy text documents [5]. Generating effective summaries of these documents can help to minimize the impact of information overload. Recently the research community within text mining has looked into various methods for dealing with this problem [1]. Keyword extraction is one strategy that summarizes important aspects of a document, while enabling the reader to quickly contrast among documents using the selected keywords [2, 16, 18, 24].

Another emerging information-seeking paradigm is one where a user's information need is vague and exploratory in nature [22]. In these cases, classification of information nuggets into various facets, commonly known as "faceted search" [20], helps users in this knowledge discovery task. Similar to keyword extraction, faceted search attempts to organize unstructured information into a structure leading to a paradigm of exploratory information seeking [20]. Traditionally, faceted search systems organize document collections into various attributes, and allow the user to navigate along the various organized attributes.

Despite sharing a common goal of organizing unstructured text into structured information, to the best of our knowledge keyword extraction and faceted search have not been explored in combination with each other. Previous approaches for faceted search

have attempted to extract various facets from within textual metadata [4][1] and tend to limit exploration to useful knowledge hidden within the textual content. Such exploration can be particularly helpful for textual data containing a variety of useful topics from which meaningful inferences can be made. Consider for example the case of a journalist wishing to examine the various ways in which news sources are reporting on different issues that are relevant to an on-going election campaign. In such a scenario, faceted search can offer a way to support extensive navigation required by the journalist, and more so if the facets are extracted from within the text appearing in news articles rather than metadata alone. This preliminary study proposes to utilize keywords for the refinement of faceted search thereby leading to exploration of novel information, and presents as a case-study extraction of keyword-based facets from within news articles. Our approach to keyword extraction models a document as a graph of terms to which we apply biased PageRank followed by maximization of topical coverage through extraction of diverse aspects of a given topic. We demonstrate the effectiveness of keyword-based facets through a system-centric evaluation on a dataset of news articles collected from eight different Irish and international news sources.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of related work. In Section 3, we present a detailed explanation of the manner in which we extract diverse keyword-based facets. In Section 4, we present experimental evaluations on various queries applied to the news article dataset to demonstrate the usefulness of our approach. In Section 5, we conclude the paper with a discussion of possible future extensions.

2 Related Work

Our work here touches on a number of different fields. In the following we review related work in keyword extraction together with faceted search. We also highlight how our work differs from existing approaches.

2.1 Keyword Extraction

Recent years have seen keyword extraction as a dominant technique for summarizing the contents of a text corpus, with numerous applications in various information access tasks such as query expansion, document classification, and document clustering, to name but a few.

Due to the differences in the nature of textual documents, generally four document specific factors have influenced keyword extraction techniques: document length, structural consistency of the document, possibility of topic change within the document, and possibility of topic correlation among topics within the document [7]. The longer the document, the more candidate keywords are available (e.g. scientific articles and technical reports compared to news articles and emails). A well structured document contains certain sections (fields) and formatting that can be exploited for keyword extraction,

[1] The most popular faceted search interfaces have been deployed on e-commerce sites where the data contains pre-defined attributes (price, genre etc.) from which facets are extracted.

such as a scientific paper's title and abstract [12], and metadata of webpages [25]. Documents such as conversational texts, logs of open-ended meetings generally contain several topics in sequence (as in talking points), and in such type of documents a topical change happens as the discussion moves [11], e.g., first topic can be about cleaning, second can be related to cooking, etc. Documents such as news articles and scientific articles can possess a topical correlation (i.e., interconnected topics) in the entire flow of the article, unlike informal chat. Therefore, in these type of documents the keywords are usually related to one another [18, 21].

Several approaches have been proposed in the literature to address the problem of keyword extraction from a piece of text. However, keyword extraction is generally performed in two steps. First a list of candidate keywords are extracted using some heuristics, and then each candidate is scored using either a supervised or an unsupervised strategy. A candidate keyword is typically extracted on the basis of n-grams [10, 17], words with specific parts of speech tags (i.e. nouns, verbs, adjectives) [14, 18], noun phrases [23], words other than stopwords [15], and n-grams appearing as Wikipedia articles titles [6]. Scoring each candidate keyword in a supervised strategy is influenced by the selection of different features and by the process of task re-definition. Scoring in an unsupervised strategy is often addressed using graph-based approaches or topic modeling.

2.2 Faceted Search

Faceted search has been dominant in commercial applications, with prime examples being e-commerce sites such as those of IBM Websphere and Amazon. Within the academic setting, Flamenco by Hearst [9] is one of earliest faceted search systems which uses rich metadata in a flexible manner to guide the user's navigational behavior. The traditional definition of facets says that it consists of a "a set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain" [8]. Earlier works however are limited in the sense that properties and dimensions of document collections are used to extract facets and textual data with its inherently unstructured nature cannot be organized properly into facets. This led the research community towards methods for automatic facet extraction through lexical subsumption [3], synsets and hypernym relations from WordNet [19], and personalized PageRank on ODP categories [13]. These efforts remain limited to defining concept hierarchies for document collections thereby limiting information discovery to a few broad concepts.

We propose to advance faceted search through the utilization of keyword-based facets, and present a prototype of such a system applied to news articles. The system aims to address limitations of current faceted search systems by facilitating navigation at a higher granularity than what is allowed by current systems. To this end, we extract keywords from within retrieved news articles in response to a certain query, while maximizing the topical coverage and hence, the diversity of extracted keywords. Note that this differs from traditional keyword extraction where the document corpus is static in nature.

3 Methodology

In this section, we present an overview of our methodology for the extraction of diverse keywords. We first explain how we utilize biased PageRank for the process of keyword extraction followed by an explanation of our technique to maximize diversity within the extracted keywords.

3.1 Keyword Extraction

As mentioned in Section 1, we apply PageRank to the graph of terms appearing in a given document. The effectiveness of graph-based ranking algorithms in Web search applications has encouraged researchers to apply similar models to textual data for natural language processing. TextRank by Mihalcea and Tarau [18] is an example of such a model, and we follow a similar intuition by treating terms within a document as nodes with edges between terms that co-occur. It is significant to note that the existing keyword extraction techniques utilise a static corpus while our computation is of a dynamic, real-time nature over a set of documents generated in response to a query[2]. We list below a number of ways in which our proposed technique differs from TextRank:

- We apply the keyword extraction model over a collection of documents instead of using a single document.
- We define edge weights between the terms with respect to word distance (as an exponential decay factor) between them, instead of uniform edge weights within a fixed window length.
- We use the relevance score of each document in relation to the query to further bias the PageRank node vector (i.e., terms extracted from a each document are biased towards the relevance score).
- Lastly, we identify significant terms from the corpus through chi-square test of independence, and further bias the PageRank node vector for these significant terms and named entities in the document collection.

We now explain the various steps of our keyword extraction process:

1. We apply cost-effective, time-series-based clustering to the ranked list of documents[3]. We cluster the retrieved documents using a single feature which is the creation time-stamp of each document. We argue that sub-topics of similar interests are usually clustered around a specific time window. For example: pre-election, on the election day, and post-election can be three different time windows, clustering various sub-topics around the main topic "elections".
2. We pick the top retrieved articles, in proportion to the size of each cluster. This reduces the full set of documents to a representative sub-sample of documents prior to the application of PageRank. For example, the volume of documents may be higher around the election dates as opposed to a few months before the elections hence resulting in a larger number of articles for the "election" cluster.

[2] Recall from Section 1 and 2 that our main goal is to propose keywords as facets for information navigation in a faceted search system.

[3] These are basically the ranked list of documents retrieved in response to a query

3. We apply biased PageRank to the reduced set of documents from the previous step and extract single terms representative of the document collection. To ensure further computational efficiency, we compute the PageRank scores for terms appearing in titles and sub-titles of documents only, which helps reduce the size of the term graph for real-time operation.

Finally, we merge single terms identified by biased PageRank to extract bigrams keywords. To achieve this, we add the individual PageRank scores of the co-occurring terms according to their probability of co-occurrence implying that n-grams that co-occur frequently within the documents retrieved are highly likely to constitute a keyword. This step utilises the terms in entire document instead of titles and sub-titles alone.

3.2 Maximizing Topical Coverage of Extracted Keywords

One potential limitation of utilizing biased PageRank for keyword extraction is the selection of keywords that are redundant in terms of coverage over a range of different sub-topics. This limits the coverage of diverse topics in response to a given query. We illustrate with an example from the news domain: in response to the query "syria" our algorithm retrieves keywords "syria state", "syria russian" and "syria government". which do not cover a wide variety of topics. It is therefore essential to ensure a coverage of a maximum range of diverse topics through the extracted keywords (or facets). We utilize the documents retrieved with a given keyword to measure the "freshness" of a given keyword. Here, freshness is the probability of the number of unique documents that a certain keyword is able to retrieve compared to other keywords (or facets), and this is then multiplied with the PageRank score calculated from Section 3.1.

Finally, keywords with highest scores after the application of proposed freshness strategy are extracted as facets. To return to our example query "syria", our algorithm is now able to retrieve keywords "syria talks", "syria un" and "syria vote" which are more diverse and maximize the coverage over a range of different sub-topics.

News Source	**No. of Articles**
Independent	50,309
BBC	35,179
Irish Times	30,919
Irish Examiner	19,258
RTE	18,299
Reuters	17,899
The Journal	12,020
Al Jazeera	2,786

Table 1. Information about news sources and articles in each news source.

Algorithm	Facets
Proposed algorithm	independent water, cork water, irish policy
TextRank	future, charges, system
TFxIDF	new ireland, ireland new, year ireland

Table 2. Top-3 facets by various algorithms for the query "Irish water".

Algorithm	Facets
Proposed algorithm	education secretary, education system, education scheme
TextRank	system, next, years
TFxIDF	school education, new school, education schools

Table 3. Top-3 facets by various algorithms for the query "education".

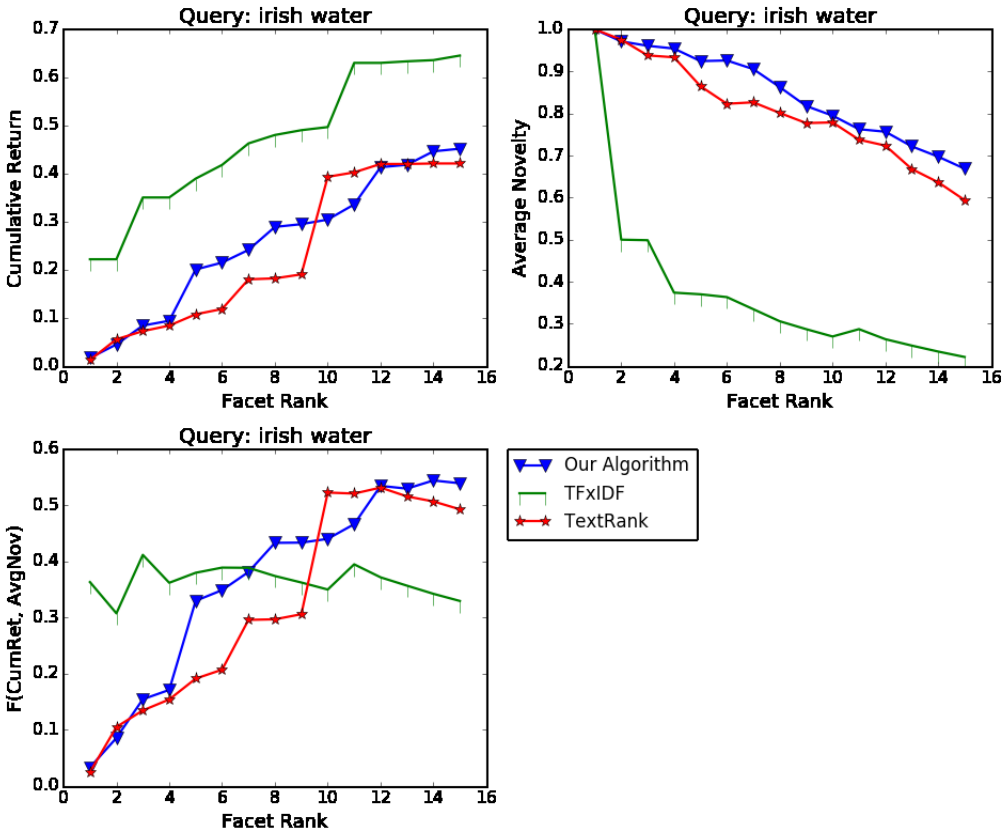

Fig. 1. Experimental results for "cumulative return", "average novelty", and "f-measure" corresponding to the query"Irish water".

Algorithm	Facets
Proposed algorithm	housing minister, housing top, cork council
TextRank	market, house, year
TFxIDF	housing dublin, dublin housing, housing crisis

Table 4. Top-3 facets by various algorithms for the query "housing".

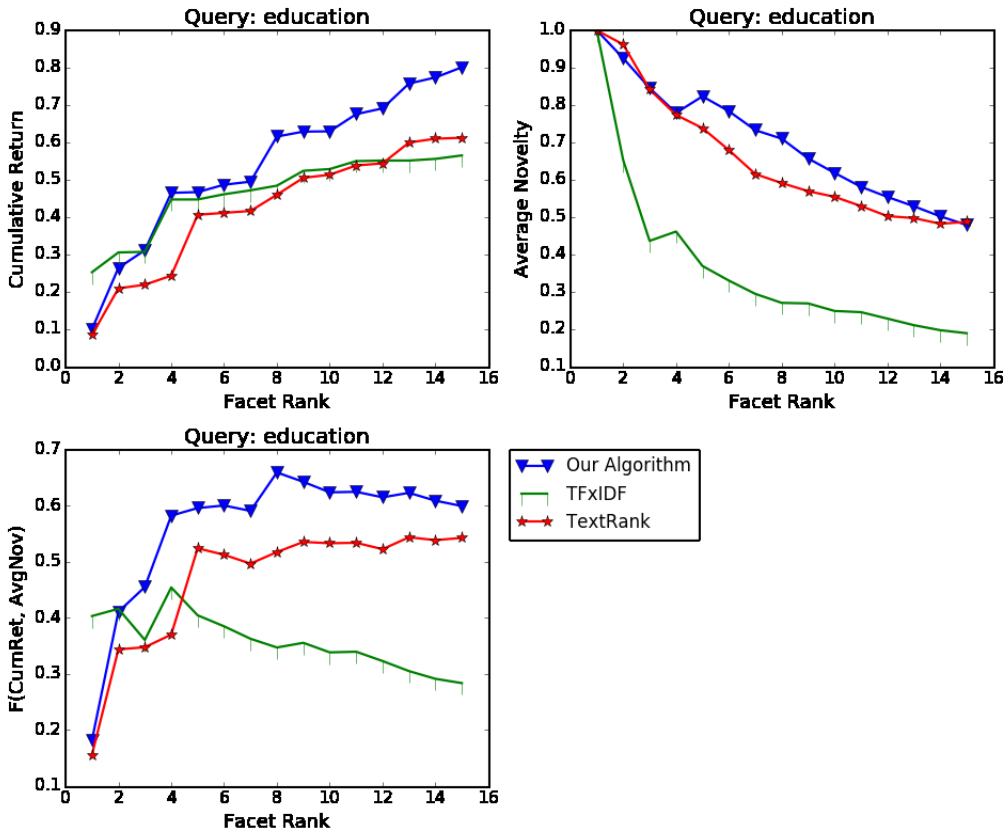

Fig. 2. Experimental results for "cumulative return", "average novelty", and "f-measure" corresponding to the query "education".

4 Experimental Evaluations

In this section, we present experimental evaluations over two system-centric measures which demonstrate the strength of our keyword-based facets in a real-time information retrieval setting. Our dataset comprises news articles extracted from eight different news sources,[4] and we show results for three different information access needs. Table 1 shows detailed statistics about the number of articles from each news source, and they cover a period from 8th July, 2015 to 7th April, 2016. We utilize system-centric

[4] The Irish Independent, The Irish Times, The Irish Examiner, RTÉ, The Journal, BBC, Reuters, Al Jazeera.

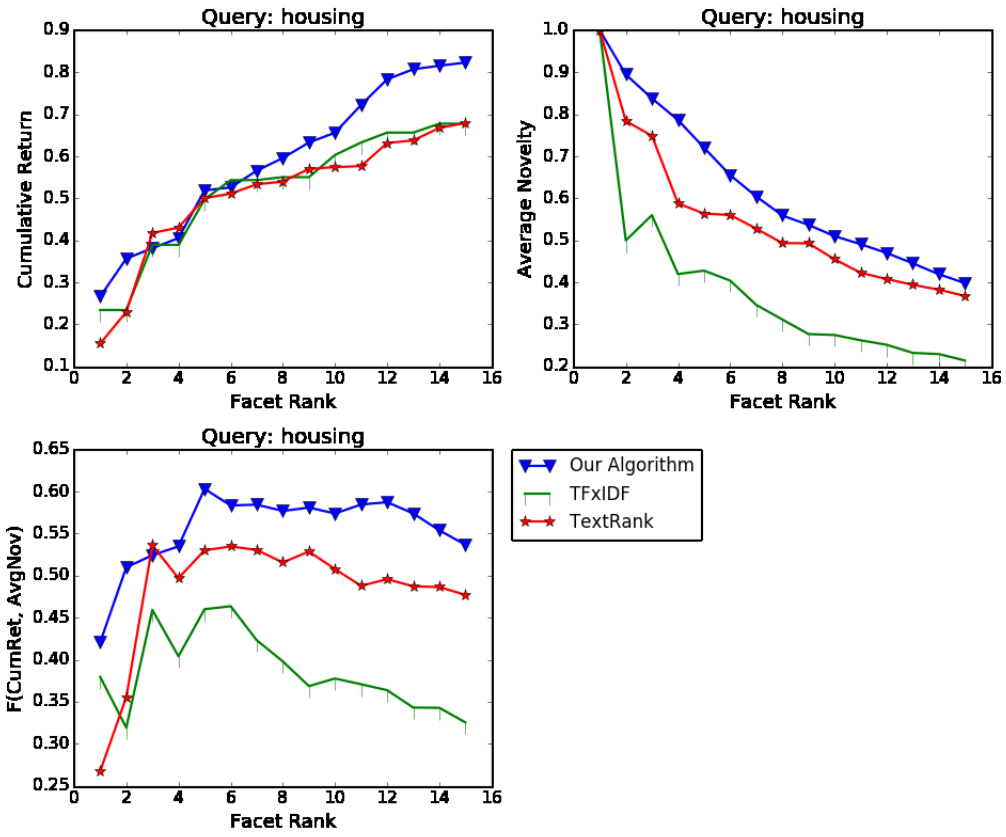

Fig. 3. Experimental results for "cumulative return", "average novelty", and "f-measure" corresponding to the query "housing'.'

evaluation measures, namely, *cumulative return* and *average novelty*. Cumulative return is defined as the coverage of the total number of documents retrieved by all facets relative to the total number of documents returned without using faceted search. Average novelty is the average ratio between the unique documents retrieved at each pair of successive facets. Both measures effectively help us to quantify the various ways in which facets aid the process of information navigation – cumulative return indicates the potential of facets to return as many documents as possible; average novelty indicates the potential of facets to return as many undiscovered documents as possible. We use these measure to compare our proposed approach to two existing competing approaches – TextRank and TFxIDF.

In what follows we present experimental results for three important issues of concern during Irish General Election 2016, namely "irish water", "education", and "housing". A journalist monitoring various issues surrounding Irish Elections 2016 needs ex-

tensive information navigation capabilities for queries issued in response to the above issues. Tables 2, 3 and 4 show the facets returned by our algorithm and the two competitors. From these results, it is evident that our algorithm extracts informative facets. As further proof of concept, we utilize the facets returned by the three algorithms to retrieve more documents and plot "cumulative return" together with "average novelty" for each of the three queries. Figures 1, 2 and 3 show the experimental results for each of these queries; and we also give an overall picture through f-score[5] between "cumulative return" and "average novelty".

In the case of the query "Irish Water" as shown in Figure 1, TFxIDF demonstrates higher "cumulative return" but our algorithm fetches new documents at the higher rate as shown by "average novelty". This implies that TFxIDF is able to retrieve overall more documents through its extracted facets, but these documents are retrieved under general facet terms which are generally not informative or representative of the original query "Irish Water" (refer to the list of top terms in Table 1, and observe the non-informative facets in terms of human intuition for TFxIDF). In the case of the queries "education" and "housing", our algorithm predominantly outperforms TFxIDF and TextRank.

5 Conclusion and Future Work

In this paper we have proposed an approach to utilize keywords for effective information navigation in faceted search systems. Current approaches to faceted search operate over metadata or hierarchical concepts, and we proposed an improvement over this in the form of keyword-based facets. Our technique made use of a graph-based modeling over terms of a document with the documents being retrieved in response to a query. We used significant terms and named entities to bias PageRank over the graph of terms followed by application of a diversity-aware strategy called "freshness". Experimental evaluations over issues pertaining to Irish General Elections 2016 showed the superiority of our technique. As future work, we aim to present more extensive experimental evaluations, by means of both system-centric and user-centric evaluation measures as applied to feedback coming from a user study.

Acknowledgments: This publication has emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. C. C. Aggarwal and C. Zhai. *Mining text data*. Springer Science & Business Media, 2012.
2. A. Csomai and R. Mihalcea. Linking educational materials to encyclopedic knowledge. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 557–559, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.
3. W. Dakka, P. G. Ipeirotis, and K. R. Wood. Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 768–775, Bremen, Germany, 2005.

[5] It is simply the harmonic mean between the two measures.

4. D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman. Dynamic faceted search for discovery-driven analysis. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 3–12, Napa Valley, California, USA, 2008.
5. R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.
6. M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 661–670, Madrid, Spain, 2009.
7. K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey. In *Proceedings of the Association for Computational Linguistics (ACL)*, ACL 2014, pages 1262–1273, Baltimore, Maryland, USA, 2014.
8. M. Hearst. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR workshop on faceted search*, pages 1–5. Seattle, WA, 2006.
9. M. A. Hearst. Next generation web search: Setting our sites. *IEEE Data Eng. Bull.*, 23(3):38–48, 2000.
10. A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Sapporo, Japan, 2003. Association for Computational Linguistics.
11. S. N. Kim and T. Baldwin. Extracting keywords from multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 199–208, Bali, Indonesia, 2012.
12. S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin. Automatic keyphrase extraction from scientific articles. *Language resources and evaluation*, 47(3):723–742, 2013.
13. C. Kohlschütter, P.-A. Chirita, and W. Nejdl. Using link analysis to identify aspects in faceted web search. In *SIGIR'2006 Faceted Search Workshop*, pages 55–59. Seattle, WA, 2006.
14. F. Liu, D. Pennell, F. Liu, and Y. Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 620–628, Boulder, Colorado, 2009.
15. Z. Liu, P. Li, Y. Zheng, and M. Sun. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 257–266, Singapore, 2009.
16. Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.
17. O. Medelyan, E. Frank, and I. H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1318–1327, Singapore, 2009.
18. R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004*, pages 404–411, Barcelona, Spain, 2004.
19. E. Stoica, M. A. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, pages 244–251, Rochester, New York, USA, 2007.
20. D. Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.
21. P. Turney. Coherent keyphrase extraction via web mining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 434–439, Acapulco, Mexico, 2003.

22. R. W. White and R. A. Roth. Exploratory search: beyond the query-response paradigm (synthesis lectures on information concepts, retrieval & services). *Morgan and Claypool Publishers*, 3, 2009.
23. Y.-f. B. Wu, Q. Li, R. S. Bot, and X. Chen. Domain-specific keyphrase extraction. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 283–284, Bremen, Germany, 2005. ACM.
24. S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu. Document concept lattice for text understanding and summarization. *Information Processing & Management*, 43(6):1643–1662, 2007.
25. W.-t. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web*, pages 213–222, Edinburgh, Scotland, UK, 2006. ACM.