Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction

Natalie Schluter
Department of Computer Science, School of Technology, Malmö University, Malmö, Sweden
natalie.schluter@mah.se

Abstract. The manner in which keywords fulfill the role of being central to a document is frustratingly still an open question. In this paper, we hope to shed some light on the essence of keywords in scientific articles and thereby motivate the graph-based approach to keyword extraction. We identify the document model captured by the text graph generated as input to a number of centrality metrics, and overview what these metrics say about keywords. In doing so, we achieve state-of-the-art results in unsupervised non-contextual single document keyword extraction.

1 Introduction and Previous Work

Researchers are far from reaching a consensus about what a *keyword extracted* from a document actually is, which may provide part of the explanation for why the keyword extraction task is such a challenge. Some characterisations of keywords are that, with respect to the document at hand, they are important or significant (Mihalcea & Tarau, 2004; Liu *et al.*, 2009a; Wan & Xiao, 2008), salient (Wan *et al.*, 2007), or (less directly) *central*, and that somehow they fulfill the role of little summaries for the document and represent the document. But how keywords fulfill these characteristics, is a question that is still wide open and building a system for a task that is not well-defined is difficult, if not unwise. However, the important role of keywords in tasks such as document indexing for search engines, summarisation, clustering and classification, make this (currently) ill-defined research necessary.

In this paper, we present a study on the essence of keywords in scientific articles and thereby motivate the graph-based approach to keyword extraction. We identify the document model captured by the text graph generated as input to a number of centrality metrics, and overview what these metrics say about keywords. In doing so, we achieve state-of-the-art results in non-contextual single document keyword extraction for the Inspec corpus, an NDCG score of 0.07578 ; we can affirm this, because the systems we compare here in terms of NDCG include re-implementations of the previous state-of-the-art.

We identify two broad types of *single document keyword extraction* (SDKE). *Contextual SDKE* makes use of the document set to which the relevant document belongs, and in which there are similar documents ; other information outside of the document set may also be used in some types of contextual SDKE. *Non-contextual SDKE* makes use of only the relevant document with no other information. The latter does not necessarily make the assumption of independence of documents in general. Non-contextual SDKE is actually important for the case of isolated documents (not part of a document set), as well as for documents for which relevant supplementary information may be non-existent or unreliable. In addition, the study of non-contextual SDKE is a study of baselines in keyword extraction : before constructing complex methods using more information, it is important to understand what we can achieve with the document alone.

This paper presents a study of graph-based unsupervised non-contextual SDKE. Recent studies by Mihalcea & Tarau (2004) (which was the original approach), Rose et al. Rose *et al.* (2010), Litvak, Last & Litvak *et al.* (2013), and Litvak & Last (2008), which also represent the state-of-the-art for the unsupervised version of this task, all use graph-based methods. The superior method (in (Rose *et al.*, 2010), rediscovered in (Litvak *et al.*, 2013)) of the three simply extracts the vertex degree ; the other methods apply more complex algorithms (PageRank (Brin & Page, 1998) is applied in (Mihalcea & Tarau, 2004; Liu *et al.*, 2010; Zhao *et al.*, 2011) and HITS (Kleinberg, 1999) is applied in (Litvak & Last, 2008)). What all three methods have in common, though they do not state that this is their explicit intention, is that they exploit measures of *centrality* of the graph. This paper aims to make these centrality measures explicit and study their appropriateness for the task. Our work will be comparable to that of Mihalcea & Tarau (2004) and Litvak *et al.* (2013), who also worked on unsupervised non-contextual single document keyword extraction for the same dataset, with the latter presenting the previous state-of-the-art. We carry out experiments on the test set from the Inspec abstract corpus (Hulth, 2003) consisting of 500 abstracts for scientific articles, along with the *uncontrolled* corresponding keywords.

We consider seven well-known measures of centrality, which, following (Borgatti & Everett, 2006) are distributed among the three categories : degree-like centrality, closeness-like centrality, and betweenness-like centrality. In doing so, we hope to either provide motivation, both conceptually (and/or mathematically) and empirically with respect to the appropriateness of these centrality measures. We also give the first non-results oriented motivation for the use of a graph representation of document that provides a basis for which these centrality measures are of interest.

We begin in Section 2 by motivating the document graph model. Then we turn our discussion to centrality measures (Section 3). Finally, we present the experimental results and discussion (Sections 4 and 5).

2 What *is* the graph ?

The graph model in previous work. In non-contextual SDKE, the main motivation for attempting to model the text of a document as a graph is for the use of some ranking algorithm over the textual units of the entire graph, based on some notion of centrality, where the relationships modelled between these units (i.e., the edges or directed edges) are co-occurrence relationships.

(Litvak & Last, 2008) and (Litvak *et al.*, 2013) posit that this reflects linguistic syntax. Certainly, specific linear order textual relations can result from linguistic syntax, to various degrees depending on the language in question, and to a large degree in English specifically. However, it is unclear why syntax alone should be the motivation for the creation of a text graph and its input into a ranking algorithm. What specifically is the role of syntax in determining the most important textual units of a document ?

(Mihalcea & Tarau, 2004), on the other hand, argue that the edges represent connections between concepts (approximated by text units), in terms of their cohesiveness for building up a conceptual context "web" in which a human would understand the text at hand ; in this context web, they explain, some units are more important than others, and they are detectible by their connections with other important concepts, interpreted as "recommendations". They also introduce the notion of a *text surfer*, which we interpret to be the *reader*, over this concept web (PageRank with the concept web as input), as the algorithm for detecting some inherent ranking of units in this web. From these notions, two questions immediately arise. How does a concept (word) recommend another word from mere co-occurrence ? What does it mean for a word to recommend another word ? For the case of the internet, it is clear how these (directed) links form immediate recommendations for given topics. However, this seems to not be as clear in the case of texts.

Sometimes the edges are directed (as in, for example, (Litvak & Last, 2008; Litvak *et al.*, 2013)) and sometimes undirected (as in, for example (Mihalcea & Tarau, 2004)). However, no motivation is given for directed-ness of the relationships based on the nature of the object modelled, over a results related argument from the experiments reported in these studies (they tried both and one type yielded best results), or from some other study (someone else tried both and one type yielded best results).

The graph model in this work : concept-building in communication. In this work, we model text as an *undirected* graph, where vertices are words appearing in the text and edges model text-linear relationships between words (i.e., that they are beside each other in the text) ; vertices representing semantically rich words (approximated here by non-stopwords) are collapsed into a single vertex if they are of the same form and part-of-speech. We follow (Mihalcea & Tarau, 2004), in considering the words of the text as approximations of concepts, but our model motivation differs in two very important respects, as we will now explain.

One essential difference is that we view text from the point of view of *synthesis* (text creation) rather than *analysis* (reading). We posit that (at least) for scientific text, discussing a topic efficiently requires concept coordination. In generating scientific text on a given topic (or given related topics), the "author" may require other concepts to regularly support the discussion (for example, definitions or explanations). We assume that the author is communicating in the most efficient manner possible, and that supporting concepts are named only when absolutely necessary. Moreover, we make the observation that in supporting or defining a concept, textual mention of a topic concept and supporting concepts should occur rather close to each other, in terms of the linear order of concepts (words) in the texts. We therefore approximate these concept support relations by co-occurrence relations, but recognise that these relations are essentially undirected : there is no clear order that should be observed between topic concepts and supporting concepts within a single sentence (or over several sentences for that matter). Note that the network is *not* the meaning of the documentation ; rather it is a *representation of its construction*. Flow through the concept network is seen as *communicative*concept-building on the part of the author for the reader.

Given this graph, we want to find the most "central" concepts/words and propose these as keywords for the document. However, several notions of centrality can now be applied. A concept can be considered important, because it

1. has first-hand access to many concepts (degree centrality),
2. is very close (and therefore is indirectly used as support) in the network to many distinct concepts (closeness centrality), or
3. is between many concepts and therefore might be often traversed (expressed) in order to build a discussion starting at one concept and ending at another (between-ness centrality).

We examine all these types of centralities using conventional deterministic measures for them, to present how the different centrality measures represent the reality of the gold keyword sets.

Generating the keyword graph. We first carry out sentence detection, tokenisation and part-of-speech tagging on the corpus, using the Stanford POS Tagger (Toutanova *et al.*, 2003). We remove all punctuation from individual sentences. However our sentences are segmented (unlike, for example in (Mihalcea & Tarau, 2004)) ; so we take sentence terminating punctuation into account (like (Litvak & Last, 2008)). We also use a stop-word list to create two different types of graphs, both of which can be constructed in time linear in the length of the document.

1. The **reduced** adjacency graph, which is constructed from the text stripped of stop-words. An edge between two words is added to the graph if these two words are adjacent in the text. Two word vertices decorated with words of the same form and (first letter of) part-of-speech are collapsed into a single vertex. This method follows that of (Liu *et al.*, 2009a). Note that we do not filter out words of a certain part-of-speech tag, unlike (Liu *et al.*, 2009b; Mihalcea & Tarau, 2004; Litvak & Last, 2008) as a preprocessing step.
2. The **full** adjacency graph, which is constructed from the full unstriped text. This time, however, word vertices that are not stop-words of the same form and part-of-speech are collapsed into a single vertex, whereas words that are stop-words are not. So here the text graphs should be seen as having two types of nodes corresponding to (1) candidate keyword parts for the document, and (2) stop words.

The point of using these two separate types of graphs is two-fold : (1) to test if we can eliminate a pre-processing step using by using different measures, and (2) to examine the behaviour of the different centrality measures on the denser (reduced) and sparser (full) graphs.

For Example (1), abstract 1939 from the Test File in the Inspec corpus, first used as an example in (Mihalcea & Tarau, 2004), we give the generated reduced adjacency text graph and full adjacency text graph in Figures 1 and 2, respectively.

(1) Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

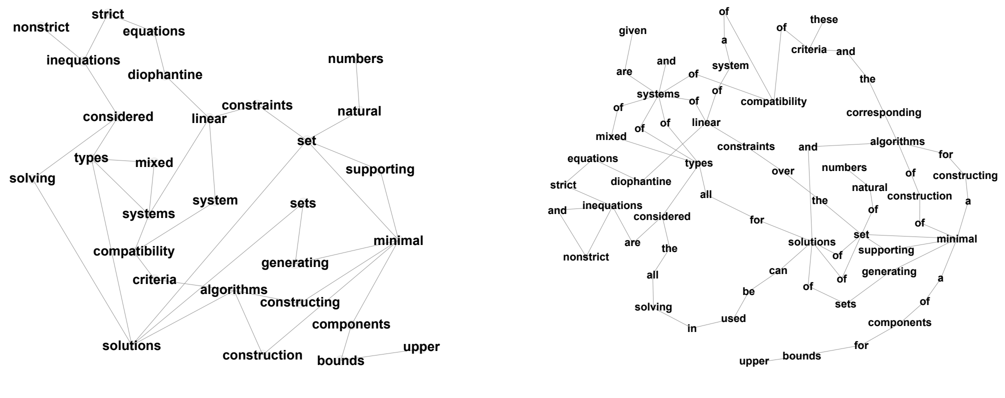

FIGURE 1 – Reduced (left) and full (right) adjacency graph for Example (1).

In this study and to maintain the *unsupervised* characterisation of the study, we use the MySQL stop-word list.[1]

3 Centrality

We hypothesise that keywords are the most "central" word vertices in a document's text graph, under some definition of centrality, which can vaguely be understood to be a summary of a node's involvement in or contribution to the cohesiveness

1. `http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html`

of the network (Borgatti & Everett, 2006). We now describe the deterministic centrality measures that we experimented with for the keyword extraction task presented in this paper.

Degree-like centrality measures. The *degree centrality* of a vertex v in a graph G, $C_D(v)$ is simply its degree $deg(v)$. Within the context of text graphs, this is a measure of how much of a first-hand support a text vertex (concept) is for other text vertices (concepts). Previous to the work presented in this paper, systems based on this measure were state-of-the-art (Litvak *et al.*, 2013; Rose *et al.*, 2010).

Eigenvector centrality is generally considered to be a measure of the “influence” of a node in a graph : the more central a node is, the more central its neighbours are and so forth. In other word, vertices are important, because they have first-hand access to many other important vertices. In the context of our concept graph approximated by the text graph, the more support a concept is provided, the more total conceptual support is offered to further concepts supported by it. The output of the PageRank algorithm is meant to be a randomised variant of this measure for *directed* graphs (Page *et al.*, 1999). In fact, this ranking is also theoretically the output of the HITS algorithm (for *directed* graphs) upon convergence (provided all eigenvalues are distinct) (Kleinberg, 1999).

To calculate the eigenvector centrality of a node $v_i \in V(G)$, $C_{EI}(v_i)$, one finds the principal eigenvector of the adjacency matrix for the graph. The ith entry in this vector is $C_{EI}(v_i)$. From this definition, it is clear that the measure cannot be used on disconnected graphs. To surmount this difficulty, we use the PageRank “teleportation trick”, transforming the input graph into a complete graph, simply incrementing the weight of all possible edges by 1.

Closeness centrality measures. *Closeness centrality* measures account for the distance of a node to all others. For computational efficiency, they consider the set of all shortest distances of a vertex x to all other nodes : $\{d(x,v) : v \in V(G)\}$. These measures are the least robust of all measures used in the following sense : for a disconnected graph, this sum is infinite for all vertices ; but we use the suggestion of (Dangalchev, 2006) for disconnected graphs, taking the limit in infinite calculations, so the distance between disconnected nodes is infinite and the reciprocal of this is just zero.

The *closeness centrality* $C_C(x)$ of a vertex x is defined as $C_C(x) := \frac{1}{\sum_{v\in V(G)} d(x,v)}$. So, the longer the distance to other word vertices (concepts), the less central the word vertex (concept) can considered. However, with this definition of closeness centrality, one cannot differentiate words that are close and far to equal numbers of nodes from those nodes that are generally close-ish to all other nodes. This is the motivation behind the *eccentricity* measure $C_{ECC}(x)$ of a vertex, which is defined as $C_{ECC}(x) := \frac{1}{\max_{v\in V(G)} d(x,v)}$.

Betweenness centrality measures. The *betweenness centrality* of a vertex quantifies how often a node acts as a bridge along the shortest path between two other nodes. In the context of our text graph, the betweenness centrality can be seen as a measure of how the presentation of a scientific subject must employ a given word (concept) as support when moving the discussion between two different concepts. We consider three different betweenness centrality measures.

The *(normalised) betweenness centrality* $C_B(x)$ for vertex x is defined as $C_B(x) := \sum_{s\in V(G)} \sum_{t\in V(G)} \frac{\sigma_{st}(x)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths between nodes s and t.

$C_B(x)$ gives more weight to pairs of vertices at a larger distance from each other. If one wishes to consider all shortest paths to contribute the same weight, one approach is to normalise by the shortest distance between s and t, which yields *length-scaled betweenness centrality*, $C_{LSB}(x) := \sum_{s\in V(G)} \sum_{t\in V(G)} \frac{\sigma_{st}(x)}{d(s,t)\sigma_{st}}$.

Finally, the *distance-weighted fragmentation* $C_{DWF}(x)$ of vertex x measures the fragmentation of a graph if we took x out of it and is defined as $C_{DWF}(x) := C_{DWF}(G-x) - C_{DWF}(G)$, where $C_{DWF}(G) := 1 - \frac{2\sum_{i\neq j} \frac{1}{d(i,j)}}{n(n-1)}$.

Note that $G-x$ (the graph obtained from G by removing vertex x and any edges incident to x) should be more fragmented than G. (We also shift all scores, so that they are positive.)

Ranked keywords. In Table 1, we give the top ranked seven words output by the degree centrality, eigenvector centrality, closeness centrality and betweenness centrality for Example (1).

$C_D(v)$	systems,	set,	minimal,	solutions,	types,	linear,	inequations
$C_{EI}(v)$	systems,	set,	minimal,	solutions,	types,	linear,	algorithms
$C_C(v)$	set,	solutions,	constraints,	minimal,	algorithms,	generating,	sets
$C_B(v)$	compatibility,	systems,	linear,	constraints,	set,	natural,	numbers

TABLE 1 – Top seven words from ranked lists for Example (1) across a selection of centrality measures, for the full text adjacency graph.

4 Evaluation

We carry out similar post-processing to (Mihalcea & Tarau, 2004). That is, sequences of adjacent keywords from the text are possibly collapsed into a multi-word keyword, depending on their scores. We score a multi-unit keyword by the average (**ave**) score of words they are composed with. This yields a candidate list where there may be unit overlaps in keywords. We therefore test an extra post-processing step which keeps only the keyword with the highest score among two overlapping keywords (this corresponds to **ave-excl** in Table 4). Ties are broken with a preference for longer keywords ; moreover, the proposed keywords must not start or end with a stop-word, and must be grounded in a noun (i.e., the rightmost word of a multi-word keyword must be a noun). Keywords consisting of at most three words are considered.

Results are evaluated using average standard Normalised Discounted Cumulative Gain (NDCG) which is mathematically proven to distinguish between ranking systems that are sufficiently different from each other (Wang *et al.*, 2013). Recall that NDCG is carried out on the *entire* ranked list and not on some top-n items ; System A is considered superior to System B if the positives (correct keywords) are generally ranked higher by System A than by System B according to the NDCG metric.

We see that there is the best performing system uses C_{LSB} (with **(ave-excl)-full**). Moreover, we observe that the post-processing step which does not allow overlapping keywords in the candidate list performs better across all measures ; this is probably because "close duplicates" that would otherwise dilute the higher ranking positions are excluded.

We also consider that the use of the set-based evaluation metrics of precision, recall, and f-score are misleading for rank-based systems, but may be informative as a means of error analysis when considering best parameter performance : poor system performance can possibly be explained by a system reaching its optimal f-score too early (for example, $n = 1$), or too late (for example, $n = 50$) in the ranked list. The best parameter f-scored system is C_B**-(ave)-full** occurs at $n = 9$, with only half of the list being true positives.

The eigenvector centrality measure is equivalent to PageRank and we test the degree centrality measure. These are the two methods that have previously been tested for non-contextual SDKE, in (Mihalcea & Tarau, 2004) and (Rose *et al.*, 2010) respectively. The length-scaled betweenness centrality measure outscores both of these measures in NDCG.

In terms of graph structure, we see that in general the full graph is preferred, which attests to the robustness of the measures and their preference for as much information as possible in the graph.

measure	post-proc	graph	n	precision	recall	f-measure	NDCG
Degree Centrality	ave	reduced	3	0.667	0.25	0.363	0.06064
		full	17	0.353	0.75	0.48	0.06115
	ave-excl	reduced	13	0.308	0.5	0.381	0.05982
		full	14	0.357	0.625	0.455	0.05830
Eigenvector Centrality	ave	reduced	4	0.5	0.25	0.333	0.06192
		full	5	0.4	0.25	0.308	0.06209
	ave-excl	reduced	4	0.5	0.25	0.333	0.05976
		full	5	0.4	0.25	0.308	0.05778
Eccentricity	ave	reduced	6	0.333	0.25	0.286	0.04264
		full	16	0.25	0.5	0.333	0.04535
	ave-excl	reduced	10	0.2	0.25	0.222	0.04769
		full	2	0.5	0.125	0.2	0.04692
Closeness Centrality	ave	reduced	6	0.333	0.25	0.286	0.04264
		full	16	0.25	0.5	0.333	0.04535
	ave-excl	reduced	10	0.2	0.25	0.222	0.04769
		full	2	0.5	0.125	0.2	0.04692
Distance Weighted Fragmentation	ave	reduced	3	0.667	0.25	0.364	0.06204
		full	6	0.5	0.375	0.429	0.06094
	ave-excl	reduced	3	0.667	0.25	0.364	0.05995
		full	6	0.5	0.375	0.429	0.05743
Betweenness Centrality	ave	reduced	3	1.0	0.375	0.545	0.06352
		full	9	0.556	0.625	**0.588**	0.06358
	ave-excl	reduced	3	1.0	0.375	0.545	0.06331
		full	3	1.0	0.375	0.545	0.06334
Length-Scaled Betweenness Centrality	ave	reduced	6	0.667	0.5	0.571	0.06660
		full	3	1	0.375	0.545	0.06658
	ave-excl	reduced	4	0.75	0.375	0.5	**0.07578**
		full	3	1	0.375	0.545	0.07550

TABLE 2 – NDCG and best parameter (n) precision, recall and f-score for all centrality measures tested.

5 Future Work

We have presented a study on the essence of keywords in scientific text, showing that firm understanding of the "flavour" of centrality we are trying to predict is essential in keyword extraction tasks. Some open questions remain. For instance, how robust are these measures on large document graphs ? Also, with larger graphs, time and space becomes an issue : how can these measures be efficiently computed in general ? Large documents pose the next challenge.

Références

BORGATTI S. P. & EVERETT M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, **28**(4), 466 – 484.

BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

DANGALCHEV C. (2006). Residual closeness in networks. *Physica A : Statistical Mechanics and its Applications*, **365**(2), 556–564.

HULTH A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.

KLEINBERG J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, **46**(5), 604–632.

LITVAK M. & LAST M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, p. 17–24, Stroudsburg, PA, USA : Association for Computational Linguistics.

LITVAK M., LAST M. & KANDEL A. (2013). Degext : a language-independent keyphrase extractor. *J. Ambient Intelligence and Humanized Computing*, **4**(3), 377–387.

LIU F., PENNELL D., LIU F. & LIU Y. (2009a). 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, p. 620–628, Stroudsburg, PA, USA : Association for Computational Linguistics.

LIU Z., HUANG W., ZHENG Y. & SUN M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 366–376, Stroudsburg, PA, USA : Association for Computational Linguistics.

LIU Z., LI P., ZHENG Y. & SUN M. (2009b). 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, p. 257–266, Stroudsburg, PA, USA : Association for Computational Linguistics.

MIHALCEA R. & TARAU P. (2004). Textrank : Bringing order into text. In *Proceedings of EMNLP*, p. 404–411.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1999). *The PageRank Citation Ranking : Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

ROSE S., ENGEL D., CRAMER N. & COWLEY W. (2010). *Automatic Keyword Extraction from Individual Documents*, In *Text Mining. Applications and Theory*, p. 1–20. John Wiley and Sons, Ltd.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Stroudsburg, PA, USA : Association for Computational Linguistics.

WAN X. & XIAO J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, p. 855–860 : AAAI Press.

WAN X., YANG J. & XIAO J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *ACL*.

WANG Y., WANG L., LI Y., HE D. & LIU T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Proceedings of COLT*, p. 25–54.

ZHAO W. X., JIANG J., HE J., SONG Y., ACHANANUPARP P., LIM E.-P. & LI X. (2011). Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 379–388, Stroudsburg, PA, USA : Association for Computational Linguistics.