Weighting of Passages in Question Answering

Vít Novotný and Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
witiko@mail.muni.cz, sojka@fi.muni.cz

Abstract. Modern text retrieval systems employ text segmentation during the indexing of documents. We show that, rather than returning the passages to the user, significant improvements are achieved on the semantic text similarity task on question answering (QA) datasets by combining all passages from a document into a single result with an aggregate similarity score. Following an analysis of the SemEval-2016 and 2017 task 3 datasets, we develop a weighted averaging operator that achieves state-of-the-art results on subtask B and can be implemented into existing search engines. Segmentation in information retrieval matters. Our results show that paying attention to important passages by using a task-specific weighting method leads to the best results on these question answering domain retrieval tasks.

Keywords: passage retrieval; question answering; Godwin's law

1 Introduction

The standard bag-of-words vector space model [17] (VSM) represents documents in terms of word frequencies as vectors in high-dimensional real inner-product spaces. The model disregards word order, which immediately limits its ability to capture the meaning of a document. Nevertheless, the inner product provides a notion of document similarity that is well-understood and scales to large datasets. As a result, the VSM forms the basis of popular inverted-index-based search engines such as Apache Lucene [2], and improvements to the VSM will have an immediate impact on the performance of many text retrieval systems.

Long documents that cover a range of different topics provide a significant challenge for the VSM, since they are difficult to statically summarize, and deemed irrelevant to most queries. For that reason, Hearst and Plaunt [7] suggested "motivated segments", segmentation that reflects the text's true underlying subtopic structure, which often spans paragraph boundaries. The method for passage retrieval that requires a NLP-parser and a semantic representation in Roget-based vectors was suggested by Prince and Labadié [16]. Keikha et al. [8] evaluated passage retrieval methods and showed that the existing methods are not effective for the passage retrieval task, and also observe that the relative performance of these methods in retrieving answers does not correspond to their performance in retrieving relevant documents. Carmel et al. [3] developed contextualisation approach for passage retrieval.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 31–40, 2018. © Tribun EU 2018

Recently, we suggested several notable improvements. Based on machine learned word vector space semantic models the indexed documents are segmented into *semantically coherent* passages, to retrieve these passages instead of the original documents. In this paper, we focus on the frequent case, when the search engine is expected to retrieve full documents rather than just the passages relevant to a query. It would seem that, in this scenario, passages are useful for the summarization of results at best. Contrary to this intuition, we show that for question answering (QA) datasets, combining the evidence of similarity provided by the retrieved passages yields significant improvements on the text similarity task compared to the VSM on unsegmented documents. Our results are fully reproducible.[1] [14]

The paper is structured as follows: In Section 2, we review the related work. In Section 3, we give an overview of our system without delving into the specifics of our datasets. In Section 4, we describe our datasets and the experimental setup. Section 5 reports and interprets the results. We conclude in Section 6 with a summary of our results, and suggestions for future research.

2 Related work

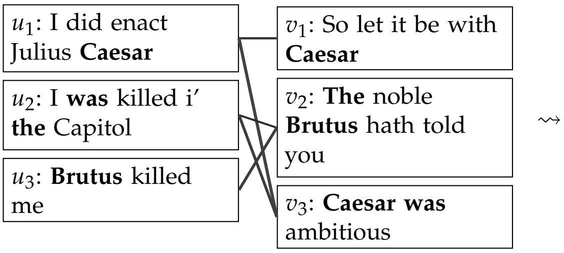
The notion of representing a document as a vector of weighted term frequencies, and estimating the similarity between two documents by the inner product was perhaps first researched by Salton and Buckley [17] during their work on the SMART information retrieval system. Several competing methods for assigning term weights and normalizing document vectors were proposed in literature. [5] In this paper, we consider those originally presented by Salton and Buckley [17].

The task of retrieving only the portions of a document that are relevant to a particular query is known as the passage retrieval and was perhaps first researched by O'Connor [15], who suggested retrieving document titles, abstracts, and figure captions in the absence of full texts. In the context of full-text retrieval, Khalid and Verberne [9] divide passage retrieval systems to those that index each passage as a separate document, which are the kind of retrieval systems that we target in this paper, and systems that first retrieve relevant documents and then retrieve passages from the retrieved documents, which is the inverse of our technique where we first retrieve passages and then aggregate the retrieved passages into documents. Beside disjoint passages, which we consider in this paper, Khalid and Verberne [9] also recognize *sliding passages* that can overlap arbitrarily.

Assessing the similarity of two structured documents by combining the evidence of similarity provided by their *structural elements* (i.e. passages) has already been explored in the context of XML document retrieval. In this paper, we draw inspiration from IBM Haifa's JuruXML system described by Mass et al. [11]. However, whereas XML documents have a tree structure, which makes it possible to compare passages based on structural similarity, our system makes no assumptions about the structure of passages.

[1] `https://github.com/witiko-masters-thesis/segmentation`

The removal and the weighting of *document zones* (i.e. passages) has been of interest to researchers in the fields of text summarization, feature selection, and text classification. In this paper, we consider the approach of Kołcz et al. [10] to reduce the number of considered passages.


u1: I did enact Julius Caesar
u2: I was killed i' the Capitol
u3: Brutus killed me
v1: So let it be with Caesar
v2: The noble Brutus hath told you
v3: Caesar was ambitious
⇝

$$\rightsquigarrow \quad \mathbf{M}_{uv} = \begin{pmatrix} 0.18 & 0 & 0.26 \\ 0 & 0.16 & 0.24 \\ 0 & 0.24 & 0 \end{pmatrix} \quad \rightsquigarrow \quad \obar_k \ominus_l m_{kl} = 0.205$$

Fig. 1. Given query and result documents u and v consisting of passages u_1, u_2, u_3, v_1, v_2, and v_3, we compute a similarity matrix $\mathbf{M}_{uv}$ using the `bnc.bnc` tf-idf weighting scheme [17]. Using the operators $\obar = \ominus = \mathrm{wavg}_{\mathrm{length}}$, we compute the aggregate score $S'(u, v)$.

3 System description

Our system takes a list of passages that form a single document and preprocesses them. If a passage k comes from a result document, then we store k in the database. If a passage k comes from a query document u, then we search the database for candidate passages l that have at least one term in common with k and we compute the similarity score $S(k, l)$. With the VSM, we first convert the passages k and l to the orthonormal coordinates of the passage vectors $\mathbf{v}_i$ and $\mathbf{v}_j$. In this paper, we perform the conversion using the `bfx.tfx` tf-idf weighting scheme suggested by Salton and Buckley [17] for short and homogeneous passages, which fits well with our QA datasets. The similarity score S between the two passages then corresponds to the inner product between $\mathbf{v}_k$ and $\mathbf{v}_l$, i.e. $S(k, l) = \langle \mathbf{v}_k, \mathbf{v}_l \rangle = \mathbf{v}_k^\top \mathbf{v}_l$.

If we performed no segmentation, then a passage corresponds directly to a document. In this scenario, we return to the user a list of candidate passages l ordered in the decreasing order of $S(k, l)$, where k is the single query passage. If we performed segmentation, then for each document v (result document) containing at least one candidate passage l, we compute a similarity matrix $\mathbf{M}_{uv}$, where every row contains the similarity scores between a single query passage from u and all passages from v (result passages) and every column contains the scores between all query passages from u and a single result passage from v. We seek an aggregate scoring function $S'(u, v)$ defined in terms of the elements of $\mathbf{M}_{uv}$, such that the ordering of result documents induced by S' correlates with the relevance of the result documents v to the information need behind the query document u.

Let $\oplus$ and $\ominus$ be weighted averaging operators on $\mathbb{R}$ and let m_{kl} denote the value of a matrix $\mathbf{M}_{uv}$ in the row and column corresponding to the query and result passages k and l. Then we can express our aggregate scoring function S' as $\oplus_{\text{query passage } k \in u} \ominus_{\text{result passage } l \in v} m_{kl}$ (see Fig. 1). In our experiments, we evaluated two operators, namely $\text{wavg}_{\text{length}}$, which assigns weights proportional to the number of tokens in a passage, and $\text{wavg}_{\text{Godwin}}$ that we will develop as a part of our dataset analysis (see Section 4.3).

When the number of passages in a document is large, the computation of $\mathbf{M}_{uv}$ can be prohibitively slow. One possible approach to speeding up the retrieval is to avoid the segmentation of query documents and to segment only the result documents instead. This is the standard approach in semi-structured XML retrieval [11], where the query constitutes only a single branch of an XML document tree. An alternative approach would be to assume that the similarity score between the query passages and the non-candidate result passages is close to zero. Instead of retrieving all results passages from v, we would fill the columns corresponding to non-candidate result passages with zeros.

4 Experimental setup

In this section, we will describe the datasets that we used for our experiments. We will then describe how we preprocessed, and analyzed the datasets.

4.1 Datasets

We evaluated our system on the SemEval-2016 and 2017 task 3 subtask B QA datasets. These datasets consist of discussion threads from the Qatar Living[2] internet forum. Given an original question, and a set of ten candidate threads, the task is to rank the candidate threads by their relevance to the original question. A candidate thread contains a related question, and the first ten comments in the thread. The performance of a system is evaluated by its mean average precision (MAP) according to the relevance judgements from the datasets. [13, 12]

The SemEval-2016 task 3 subtask B datasets consist of a training dataset (267 original questions, 1,790 threads), a dev dataset (50 original questions, 244 unique threads), and a test dataset (70 original questions, 327 unique threads). The winning SemEval-2016 task 3 subtask B submission was from UH-PRHLT-*primary* [6] with a MAP score of 76.70 who ranked threads using support vector machines (SVMs), and crafted features. The SemEval-2016 task 3 subtask B information retrieval (IR) baseline had a MAP score of 74.75.

SemEval-2017 task 3 subtask B uses the same training and dev datasets as SemEval-2016 with the provision that the SemEval-2016 test dataset can be used

[2] `http://www.qatarliving.com/forum`

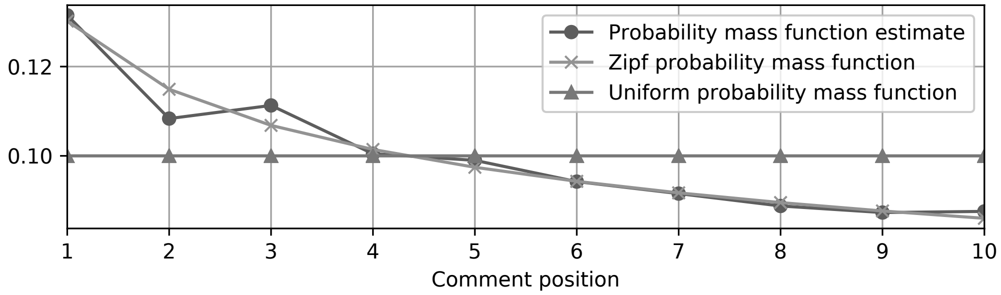


Fig. 2. Probability mass function (PMF) estimate $\hat{P}(\text{at position } i \mid \text{relevant})$ plotted along the PMF of the Zipf distribution with parameters $n = 10$ and $s = 0.18$. If the position of a comment and its relevance were independent, we would expect the PMF estimate to be uniformly distributed. Since $P(\text{at position } i)$ is uniformly distributed, $P(\text{at position } i \mid \text{relevant})$ is proportional to $P(\text{relevant} \mid \text{at position } i)$.

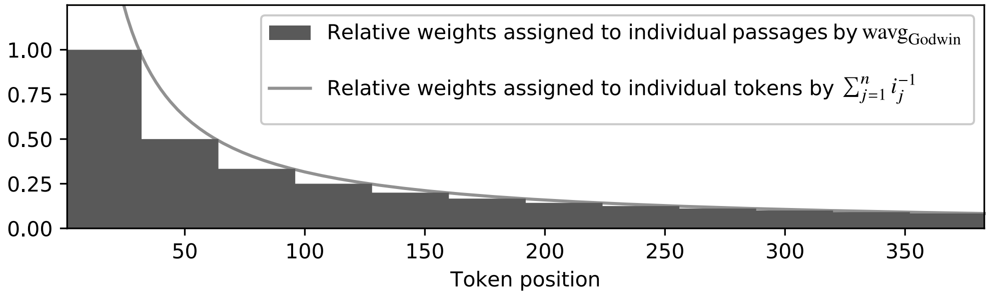


Fig. 3. The relative weights assigned to the individual passages by the $\text{wavg}_{\text{Godwin}}$ weighted averaging operator and the relative weights assigned to the individual tokens by the Godwin term weighting method. The figure assumes the mean number of tokens per a thread in the subtask A unannotated datasets (383 tokens), a uniform number of tokens in a passage, and the `txx.txx` tf-idf weighting scheme.

for training. A new test dataset (88 original questions, 293 unique threads) has also been added. The SemEval-2017 task 3 subtask B winning configuration was *SimBow-primary* [4] with a MAP score of 47.22 who ranked threads using logistic regression and unsupervised similarity measures. The SemEval-2016 task 3 subtask B IR baseline had a MAP score of 41.85.

For statistical analysis, we used the SemEval-2016, and 2017 task 3 subtask A datasets. These datasets contain equivalent data as the subtask B training datasets (2,654 questions), but now the relevance judgements assess how relevant a comment is to a question. For language modeling, we used the unannotated SemEval-2016 and 2017 task 3 subtask A datasets (189,941 questions, 1,894,456 comments).

4.2 Language modeling, and segmentation

Texts in datasets were lower-cased, stripped of images and URLs, and tokenized on white spaces and punctuation. Tokens shorter than two characters or longer than 15 characters were removed to cope with the problem of missing and extra whitespaces in questions, and comments. Using the existing structure of the datasets, every original question was split into two passages corresponding to the question subject and text, and every candidate thread was split into twelve passages corresponding to the related question subject and text, and the initial ten comments.

Since the annotated datasets did not contain enough text to build a proper language model, we used the unannotated subtask A datasets to obtain the collection-wide statistics required to compute the scoring function S described in Section 3.

4.3 Dataset analysis

In 1991, the American attorney and author Mike Godwin formulated[3] a rule that "as a Usenet discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches one." An immediate corollary would be that as an online discussion grows longer, the probability of a relevant contribution approaches zero. We were curious whether the datasets would confirm these observations. We used the subtask A relevance judgements to estimate the probability mass function (PMF) $P(\text{at position } i \mid \text{relevant})$ for $i = 1, 2, \ldots, 10$. Since there is a uniform number of comments at each position i, i.e. $P(\text{at position } i) = 0.1$, we would expect $P(\text{at position } i \mid \text{relevant})$ to be also uniform if the position of a comment and its relevance are independent. We show in Fig. 2 that this expectation is implausible, and that there appears to be an inverse relationship between the position of a comment and its relevance.

To see if this relationship was statistically significant, we modeled the number of relevant comments at each position i as a binomial random variable $X_i \sim \text{Bi}(n, \theta_i)$ with a known number of trials $n = 2{,}410$, and an unknown probability of success θ_i. We then used the one-tailed Fisher's exact test to reject the following system of null hypotheses at 5% significance:

$$H_0^{(ij)} : \theta_i = \theta_j, \text{ where } i, j = 1, 2, \ldots, 10, \ \ i < j$$

We rejected $H_0^{(ij)}$ for any $j - i > 3$. We failed to reject $H_0^{(ij)}$ for $(i, j) = (2, 3)$, $(4, 5)$, $(5, 6)$, $(6, 7)$, $(7, 8)$, $(7, 9)$, $(7, 10)$, $(8, 9)$, $(8, 10)$, and $(9, 10)$. We used the procedure of Benjamini and Hochberg [1] to control the false discovery rate due to multiple testing.

This discovery led us to develop the $\text{wavg}_{\text{Godwin}}$ weighted averaging operator, which assigns a weight proportional to i^{-1} to a passage at position i in accordance to *Zipf's law*. This decreases the effect of comments that are likely to

[3] `news:1991Aug18.215029.19421@eff.org`

Table 1. Results for the four evaluated configurations (one primary, and three contrastive) on the SemEval-2016 task 3 subtask B test dataset. The primary configuration is highlighted in bold, whereas the winning SemEval-2016 task 3 subtask B submission and the IR baseline are highlighted in italics.

Configuration	Segm.	Text summ.	S. f. S	Aggregate s. f. S'	MAP
Primary	**Yes**		`bfx.tfx`	$\obar = \mathbf{wavg}_{\mathbf{length}}$, $\ominus = \mathbf{wavg}_{\mathbf{Godwin}}$	**76.77**
*SemEval-2016 task 3 subtask B winner (*UH-PRHLT*-primary)*					*76.70*
Third contrastive	No	FirstTwoPara	`bfx.tfx`		75.21
SemEval-2016 task 3 subtask B IR *baseline*					*74.75*
First contrastive	No		`bfx.tfx`		73.94
Second contrastive	No		`bfx.tfx`, Godwin		70.28

be irrelevant. Under the hypothesis that relevant comments are more likely to contain important terms that describe the meaning of a document, this operator pays attention to scores between those passages that are likely to contain important terms.

Since term weighting is conceptually and computationally simpler than segmentation and result aggregation, we wanted to verify that the segmentation is meaningful and that the relevance loss occurs at passage boundaries rather than at term boundaries. For that reason, we developed the *Godwin* term weighting method for the VSM scoring function S. For each term t at positions $i_1, i_2, \ldots, i_n$ in a document, the method multiplies the term frequency term-weighting component [17] with a weight proportional to $\sum_{j=1}^{n} i_j^{-1}$. It is easy to show that, given the right choice of the term frequency component (`t`) and the collection frequency component (`x`), the scoring function S induces the same ordering on unsegmented threads as the aggregate scoring function S' with $\obar = \mathrm{wavg}_{\mathrm{length}}, \ominus = \mathrm{wavg}_{\mathrm{Godwin}}$ would if the threads were segmented to one passage per a token (see Fig. 3).

5 Results

The results for the four evaluated configurations are shown in Table 1 and Table 2. The primary configuration performs segmentation with the $\obar = \mathrm{wavg}_{\mathrm{length}}, \ominus = \mathrm{wavg}_{\mathrm{Godwin}}$ operators and consistently outperforms the winning SemEval task 3 subtask B submissions. This shows that the $\mathrm{wavg}_{\mathrm{Godwin}}$ weighted averaging operator works well with our datasets and hopefully with QA datasets in general.

The three contrastive configurations do not perform segmentation. The first configuration corresponds to the base system with no extra preprocessing or weighting and is consistently outperformed by the remaining configurations as well as by the SemEval task 3 subtask B IR baselines. The second configuration

Table 2. Results for the four evaluated configurations (one primary, and three contrastive) on the SemEval-2017 task 3 subtask B test dataset. The primary configuration is highlighted in bold, whereas the winning SemEval-2017 task 3 subtask B submission and the IR baseline are highlighted in italics.

Configuration	Segm.	Text summ.	S. f. S	Aggregate s. f. S'	MAP
Primary	**Yes**		`bfx.tfx`	$⦶ = \mathbf{wavg}_{\mathbf{length}'}$ $\ominus = \mathbf{wavg}_{\mathbf{Godwin}}$	**47.45**
SemEval-2017 task 3 subtask B winner (SimBow-primary)					*47.22*
Third contrastive	No	FirstTwoPara	`bfx.tfx`		44.67
SemEval-2017 task 3 subtask B IR baseline					*41.85*
Second contrastive	No		`bfx.tfx`, Godwin		37.18
First contrastive	No		`bfx.tfx`		36.82

uses the Godwin term weighting method developed in Section 4.3 and performs on-par with the first contrastive configuration. This shows that the segmentation to semantically coherent passages is meaningful and cannot be replaced with simple term weighting. The third configuration uses the *FirstTwoPara* text summarization technique [10], which reduces a thread to the question subject, the question text, and the first comment, and outperforms all the remaining contrastive configurations as well as the SemEval task 3 subtask B IR baselines. This shows that removing all but the first comment improves the signal-to-noise ratio, but at the cost of losing important terms.

6 Conclusion and future work

Segmentation matters and so does careful weighting. By combining both, we were able to achieve state-of-the-art results on the SemEval-2016 and 2017 task 3 subtask B QA datasets using the standard bag-of-words vector space model without any semantic modeling. Our technique can be readily implemented into existing inverted-index-based search engines.

We have shown that there exists a statistically significant relationship between the position of a comment and its relevance in the SemEval-2016 and 2017 subtask A datasets. Investigating whether such a relationship exists in other QA datasets and other datasets in general will provide us with new insights to the dynamics of online discourse and lead to more effective retrieval systems.

In this paper, we assumed that passages were disjoint. This is not true in general and future research should extend our technique to sliding passages [9] that can overlap arbitrarily.

Acknowledgments. TAČR Omega grant TD03000295 is gratefully acknowledged.

References

[1] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) pp. 289–300 (1995), https://www.jstor.org/stable/2346101

[2] Białecki, A., Muir, R., Ingersoll, G., Imagination, L.: Apache Lucene 4. In: SIGIR 2012 workshop on open source information retrieval. p. 17 (2012)

[3] Carmel, D., Shtok, A., Kurland, O.: Position-based Contextualization for Passage Retrieval. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. pp. 1241–1244. CIKM '13, ACM, New York, NY, USA (2013), https://doi.acm.org/10.1145/2505515.2507865

[4] Charlet, D., Damnati, G.: SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 315–319. Association for Computational Linguistics (2017), https://www.aclweb.org/anthology/S17-2051

[5] Chisholm, E., Kolda, T.G.: New Term Weighting Formulas for the Vector Space Method in Information Retrieval. Tech. rep., Computer Science and Mathematics Division, Oak Ridge National Laboratory, Tennessee, United States (Mar 1999), https://doi.acm.org/10.2172/5698

[6] Franco-Salvador, M., Kar, S., Solorio, T., Rosso, P.: UH-PRHLT at SemEval-2016 Task 3: Combining Lexical and Semantic-based Features for Community Question Answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 814–821. Association for Computational Linguistics (2016), https://www.aclweb.org/anthology/S16-1126

[7] Hearst, M.A., Plaunt, C.: Subtopic Structuring for Full-length Document Access. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 59–68. SIGIR '93, ACM, New York, NY, USA (1993), https://doi.acm.org/10.1145/160688.160695

[8] Keikha, M., Park, J.H., Croft, W.B., Sanderson, M.: Retrieving Passages and Finding Answers. In: Proceedings of the 2014 Australasian Document Computing Symposium. pp. 81:81–81:84. ADCS '14, ACM, New York, NY, USA (2014), https://doi.acm.org/10.1145/2682862.2682877

[9] Khalid, M.A., Verberne, S.: Passage Retrieval for Question Answering using Sliding Windows. In: COLING 2008: Proceedings of the 2nd Workshop on Information Retrieval for Question Answering. pp. 26–33. Association for Computational Linguistics (2008), https://aclweb.org/anthology/W08-1804

[10] Kołcz, A., Prabakarmurthi, V., Kalita, J.: Summarization as feature selection for text categorization. In: Proceedings of the ACM CIKM Conference. pp. 365–370. ACM (2000)

[11] Mass, Y., Mandelbrod, M., Amitay, E., Carmel, D., Maarek, Y.S., Soffer, A.: JuruXML – an XML retrieval system at INEX'02. In: INEX Workshop. pp. 73–80 (2002)

[12] Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: SemEval-2017 task 3: Community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 27–48. SemEval '17, ACL, Vancouver, Canada (Aug 2017)

[13] Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A.A., Glass, J., Randeree, B.: SemEval-2016 task 3: Community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation. SemEval '16, ACL, San Diego, USA (Jun 2016)

[14] Novotný, V.: Vector Space Representations in Information Retrieval. Master's thesis supervised by Petr Sojka, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2018), https://github.com/witiko-masters-thesis/thesis

[15] O'Connor, J.: Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching. Information Processing & Management 11(5–7), 155–164 (1975)

[16] Prince, V., Labadié, A.: Text Segmentation Based on Document Understanding for Information Retrieval. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) Natural Language Processing and Information Systems. pp. 295–304. Springer, Berlin, Heidelberg (2007), https://doi.acm.org/10.1007/978-3-540-73351-5_26

[17] Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24, 513–523 (1988)