SKEEFT : indexing method taking into account the structure of the document

Pascal Cuxac, Jean-Charles Lamirel, Nicolas Kieffer

SKEEFT: indexing method taking into account the structure of the document

Pascal Cuxac[1], Nicolas Kieffer[1], Jean-Charles Lamirel[2]

[1] INIST-CNRS, 2 allée du parc de Brabois, 54519 Vandoeuvre-lès-Nancy, FRANCE

pascal.cuxac@inist.fr ; kieffer.nicolas.pro@gmail.com

[2] SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, FRANCE

jean-charles.lamirel@loria.fr

Abstract: We present Skeeft, a method that improves term extraction, taking into account the structure of the document. We consider a document as a classification, each part representing a class; a feature selection method is then used to select terms specific to each part.
The analysis of the full text thus allows the extraction of relevant terms and a better weighting.

Keywords: Term extraction, XML, Sructured document, Feature selection, Full-text.

1- Introduction

Indexing a document has multiple interests: responding to a request, making a classification, comparing texts... (Gupta and Lehal, 2017).

The main problem is not to extract terms from a document but to select those that are most relevant, i. e. those that are most representative of the content of the text analyzed. Some methods such as Termsuite (Cram and Daille, 2016) or Termostat (Drouin, 2003), use a reference corpus (often a corpus of press articles) to weight the extracted terms and then filter them. The constitution of the reference corpus is therefore essential, and it is difficult to conceive that the same corpus is optimal for processing data from different disciplines such as nuclear physics or political science.

We present Skeeft, an automatic method of indexing in free language, without resources, applied to full text (scientific articles). This choice makes it possible to move away from the scientific field and treat a multi-thematic corpus without difficulty. The analysis of the full text makes it possible to extract relevant terms and improve weighting.

The document structure is often used to target extraction areas or weight terms (You et al., 2013). Our approach is very different because we do not prioritize the different parts but we put them in competition. This simplifies processing because it is sufficient to identify the parts of the document without having to identify their "role" (introduction, methodology...).

Our hypothesis is that we do not necessarily need an external corpus to process scores of importance for words and expressions (for example TF-IDF) and thus to be able to extract the most relevant words.

2- Methodology

A document is structured in distinct parts, each centred on a specific aspect of the study (state of the art, methodology, experimentation, results, perspectives...). Potentially, the language elements used can therefore be very different from one party to another. Our goal is to use these specificities related to the parts of the document to extract the most accurate information as possible.

For this purpose, we consider a document as a classification whose parts would be classes of terms.

Thanks to a features selection method, we will compare terms extracted from the different sections of the document, then after selection we will reweight the selected terms which, after automatic filtering, will give a representation of the text

On that basis, the process takes place in 4 main stages (figure 1):

- extraction of terms for each identified part (any method can be used here);
- application of a selection of variables: a term is a variable, the part where it appears is its class of belonging (Lamirel et al., 2015);
- weighting of the selected terms for each part;
- final filtering, merging of results and display.

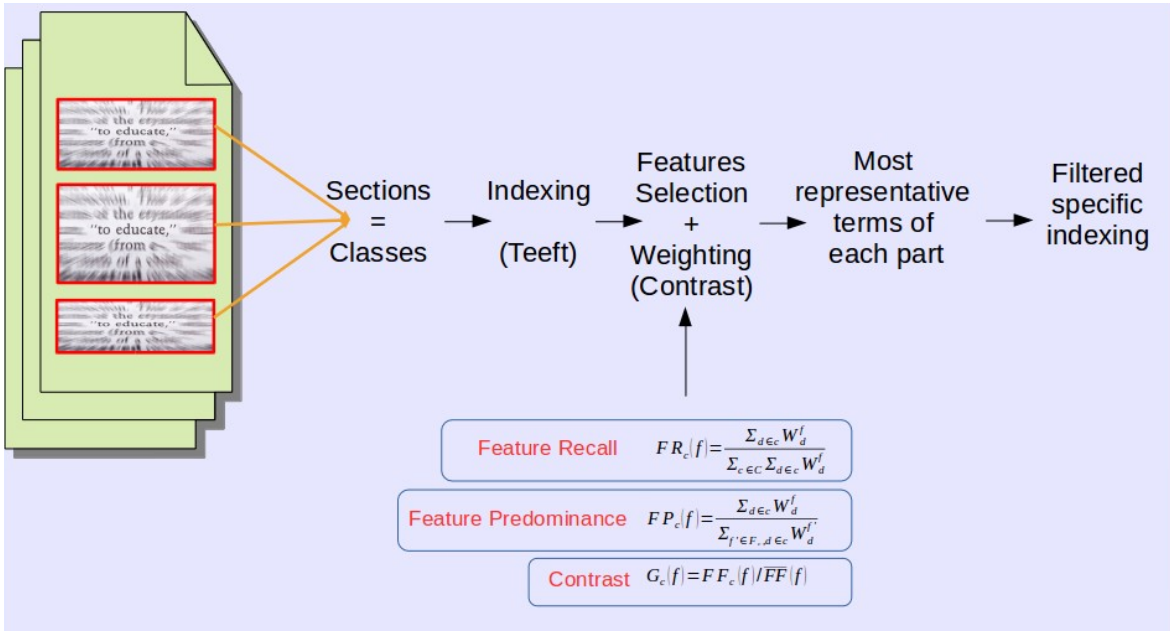

Figure 1 : General diagram

We use the term extraction tool TEEFT, a method POS we have developed to process english full-texts (Cuxac, 2017). Recently we have adapted TEEFT to treat french literature. It is interesting to note that potentially any term extraction method is usable; this means that our

approach is applicable regardless of the documents language, as long as we have a suitable term extraction method.

We apply the variable selection method based on labelling maximization developed by Lamirel et al (2015).

Let us consider a set of clusters C resulting from a clustering method applied on a set of data D represented with a set of descriptive features F, feature maximization is a metric which favors clusters with maximum Feature F-measure. The Feature F-measure $FF_c(f)$ of a feature f associated to a cluster c is defined as the harmonic mean of Feature Recall $FR_c(f)$ and Feature Precision $FP_c(f)$ indexes which in turn are defined as:

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c \in C} \sum_{d \in c} W_d^f}$$

$$FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f \in F^c, d \in c} W_d^f}$$

where W_d^f represents the weight of the feature f for data d and Fc represent the set of features occurring in the data associated to the cluster c.

Taking into consideration the basic definition of feature maximization metric, the feature maximization-based selection process can thus be defined as a non-parametric class-based process in which a class feature is characterized using both its capacity to discriminate a given class from the others ($FP_c(f)$ index) and its capacity to accurately represent the class data ($FR_c(f)$ index). The set Sc of features that are characteristic of a given class c belonging to an overall class set C results in:

$$S_c = \{f \in F_c \mid FF_c(f) > \overline{FF}(f) \; et \; FF_c(f) > \overline{FF}_D\}$$

Where $\overline{FF}(f) = \sum_{c' \in C} FF_{c'}(f) / |C_{/f}|$

And $\overline{FF}_D = \sum_{f \in F} \overline{FF}(f) / |F|$

with $C_{/f}$ C/f representing the restriction of the set C to the classes in which the feature f is present.

Features that are judged relevant for a given class are the features whose representation is altogether better than their average representation in all the classes including those features

and better than the average representation of all the features, as regard to the feature F-measure metric.

In the specific framework of the feature maximization process, a contrast enhancement step can be exploited complementary to the former feature selection step. The role of this step is to fit the description of each data to the specific characteristics of its associated class which have been formerly highlighted by the feature selection step. In the case of our metric, it consists in modifying the weighting scheme of the data specifically to each class by taking into consideration the "information gain" provided by the Feature F-measures of the features, locally to that class. For a feature c belonging to the set of selected features SC of a class C, the gain CFc can be defined as:

$$CF_c = (FF_c(f)/\overline{FF}(f))$$

The reader should refer to the article by Lamirel et al (2015) for more details.

3- Application

In our application, the starting point is the document in xml format, which is necessary to extract the intrinsic structure of the document. In the case of pdf documents, the use of a conversion tool to xml is necessary, such as PDFX[1] (Constantin et al. 2013) or Grobid[2] (Romary & Lopez, 2015).

However, we tested our approach in the event that XML was not available or in the case of a poorly structured XML. The idea is then to take the paragraphs (easily identifiable by line breaks) or even to cut the document into pages. If this is working and can give an analyzable result, the quality obtained is of course degraded. In the rest of this paper we will only discuss the case of a well-structured document.

After choosing the level of structure to use (section, subsection, paragraph...), the process will extract the most representative terms from this structure and weight them.

We use a term extraction method that we have developed (Teeft) to process scientific documents in full text. It is a PoS method that uses a generalist resource (Brown corpus) to extract relevant terms.

We then have the equivalent of a graph with nodes representing the classes (for example, sections of the document) and others representing the extracted terms, the class term links being weighted by the contrast of the terms in the class in question (Figure 2).

1 http://pdfx.cs.man.ac.uk/

2 http://cloud.science-miner.com/grobid/

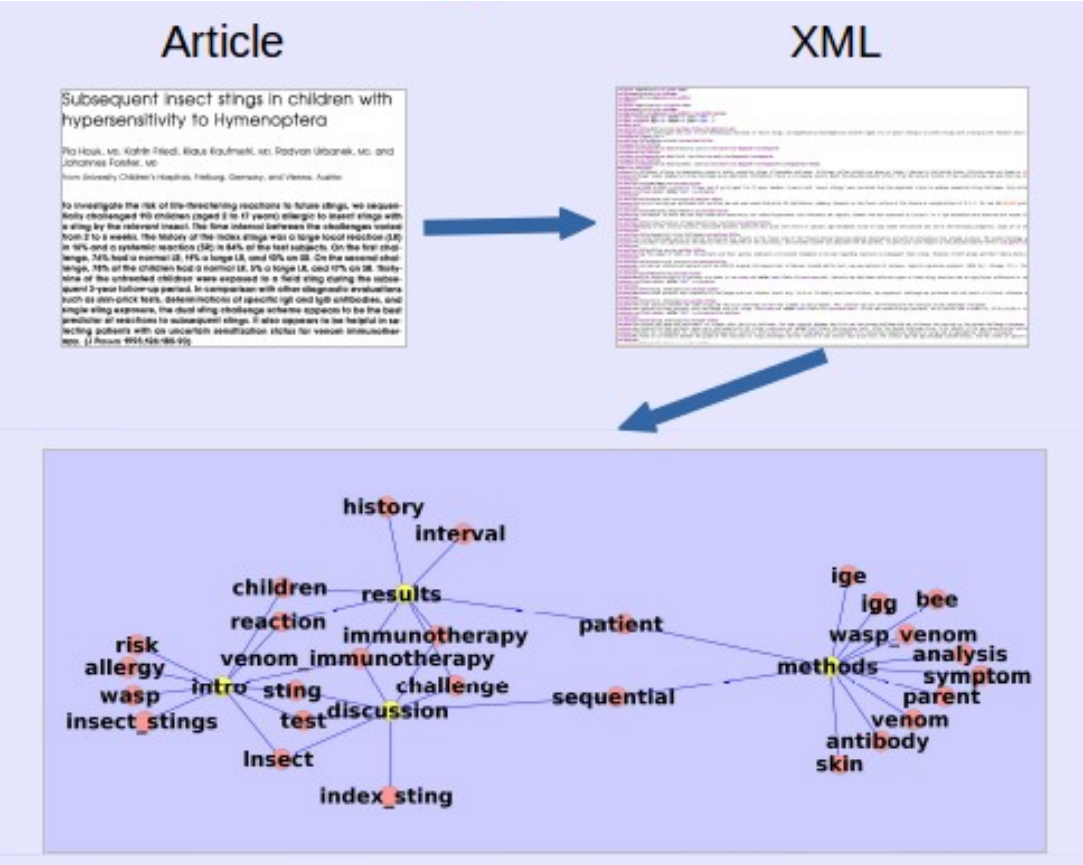

Figure 2: Graphical representation of the content of a scientific paper with the use of its structure

On this figure we see an example from an article entitled "Subsequent insect stings in children with hypersensitivity to hymenoptera" and concerning "allergic reactions of children in response to insect bites". The yellow nodes in the graph represent the sections of the document, the orange nodes represent the terms selected by our method and the edges represent the links between the terms and the sections. The selected terms, such as "Venom Immunotherapy", "Children", "Insect Stings", "Allergy", "Ige", "Igg"...are validated by an expert as characteristics of the document content.

As we will see later, this graphical representation paves the way for an automatic sentence extraction summary method.

4- Experimental results

First, Skeeft is tested on multi-disciplinary documents from the ISTEX database[3], manually indexed by experts and compared to the following methods: TopicRank (Bougouin and Boudin, 2014), Keyterm (Lopez and Romary, 2010), KPMiner (El-Beltagy and Rafea, 2009), SingleRank (Wan and Xiao, 2008), Kea[4] (Witten et al. 2000) and Termostat[5] (Drouin, 2003). We also compare with our Teeft method which does not take into account the structure of the document.

3 https://www.istex.fr/
4 http://www.nzdl.org/Kea
5 http://termostat.ling.umontreal.ca/

Table 1 - Comparison of Skeeft's performance on a manually indexed test corpus.

Méthode	Rappel	Précision	F-mesure
TopicRank	0.11	0.18	0.14
Keyterm	**0.25**	0.21	0.23
SingleRank	0.06	0.09	0.07
Kea	0.14	0.21	0.17
KPMiner	0.17	0.22	0.19
Termostat	0.24	0.30	**0.27**
Teeft	0.23	0.20	0.21
Skeeft	0.21	**0.32**	**0.25**

The results presented in the table above show good performance of Skeeft in terms of F-measurement since only Termostat gives slightly better results. It is interesting to note that compared to our traditional "Teeft" extraction method, the improvement made is significant.

In a second step we used the Nguyen2007 corpus (Nguyen & Kan, 2007). Nguyen and Kan collected a corpus containing 120 computer articles, each 4 to 12 pages long. This corpus was used for the SemEval-2010 campaign (Kim et al., 2010).

As we see in Table 2, Skeeft gives the best results. Termostat could not be applied to this data because the web interface does not accept to process corpora. Keyterm, for unknown reasons, could not analyze this corpus.

Table 2 - Comparison of Skeeft's performance on the 2007 Nguyen corpus.

Méthode	Rappel	Précision	F-mesure
TopicRank	0.12	0.10	0.10
Keyterm	-	-	-
SingleRank	0.05	0.04	0.04
Kea	0.13	0.10	0.11
KPMiner	**0.14**	0.11	0.12
Termostat	-	-	-
Skeeft	**0.14**	**0.13**	**0.14**

5- Conclusion and prospects

The Skeeft method shows good results on the two test corpora used, however we will need to find larger corpora to better evaluate our approach.

Skeeft tested here, uses the Teeft extraction method we developed. As we have seen, our approach can be adapted to any extraction method (and therefore to any language !): it will be interesting to test this methodology applied to the art methods selected here.

Skeeft can be downloaded from https://github.com/NicolasKieffer/tdm-skeeft.

Skeeft has recently been adapted to generate automatic text summaries by sentence extraction. For this, we use the selected terms in each part and their weights (contrast). This allows us to weight the sentences and then, via an automatic threshold calculation, we select the most relevant sentences. It is then just necessary to re-order the selected sentences to build the summary.

Acknowledgements:
This work was funded by the ISTEX project - ANR-10-IDEX-0004-12 program.

REFERENCES

Bougouin, A. et F. Boudin (2014). Topicrank : ordonnancement de sujets pour l'extraction automatique de termes-clés. TAL 55(5), 45–69.

Constantin, Alexandru, Steve Pettifer, et Andrei Voronkov. 2013. « PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature ». In *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13, New York, NY, USA: ACM, 177–180. http://doi.acm.org/10.1145/2494266.2494271 .

Cram, Damien, et Béatrice Daille. (2016) Terminology Extraction with Term Variant Detection. In *Proceedings of ACL-2016 System Demonstrations*, 13–18. Berlin, Germany: Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/P16-4003.

Cuxac, Pascal. (2017). ISTEX : Les projets d'enrichissement menés par les équipes de l'INIST. 10.13140/RG.2.2.30707.53281.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. Terminology 9(1), 99–117.

El-Beltagy, S. et A. Rafea (2009). Kp-miner : a keyphrase extraction system for english and arabic documents. Inf Syst 34(1), 132–144.

Gupta, V. et S. Lehal (2017). Keyword extraction :a review. Int.j.eng.appl.sci. 2(4), 215–220.

Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation*, 21–26.

Lamirel, J.-C., P. Cuxac, A. Chivukula, et K. Hajlaoui (2015). Optimizing text classification through feature selection based on quality metric. J. of Intel. Inf. Syst. 45(3), 379–396.

Lopez, P. et L. Romary (2010). Humb : Automatic key term extraction from scientific articles in grobid. In Proc. of the 5th Int. workshop on semantic evaluation, ACL, 248–251.

Nguyen, T. D., & Kan, M.-Y. (2007). Keyphrase Extraction in Scientific Publications. *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, 317–326. http://dl.acm.org/citation.cfm?id=1780653.1780707

Romary, Laurent, et Patrice Lopez. 2015. « GROBID - Information Extraction from Scientific Publications ». https://hal.inria.fr/hal-01673305 (21 juin 2019).

Wan, X. et J. Xiao (2008). Single document keyphrase extraction using neighborhood knowledge. Proceedings of the 23rd AAAI Conference on Artificial Intelligence, 855–860.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2000). *KEA: Practical automatic keyphrase extraction* [Working Paper]. Consulté à l'adresse University of

Waikato, Department of Computer Science website: https://researchcommons.waikato.ac.nz/handle/10289/1021

You, W., D. Fontaine, et J.-P. Barthès (2013). An automatic keyphrase extraction system for scientific documents. Knowledge and Information Systems 34(3), 691–724.