

Plagiarism Detection through Vector Space Models Applied to a Digital Library

Radim Řehůřek

Faculty of Informatics, Masaryk University
xrehurek@fi.muni.cz

Abstract. Plagiarism is an increasing problem in the digital world. The sheer amount of digital data calls for automation of plagiarism discovery. In this paper we evaluate an Information Retrieval approach of dealing with plagiarism through Vector Spaces. This will allow us to detect similarities that are not result of naive copy&paste. We also consider the extension of Vector Spaces where input documents are analyzed for term co-occurrence, allowing us to introduce some semantics into our approach beyond mere word matching. The approach is evaluated on a real-world collection of mathematical documents as part of the DML-CZ project.

1 Introduction

1.1 What is plagiarism?

With the advances in technology (storage and processor performance, database and scanning systems, user interfaces), creating large digital collections of documents becomes largely an engineering task. *Digital library* is a centrally managed digital collection of documents, such as texts, multimedia images, music or videos. At the same time, the electronic medium makes it easier than ever to plagiarize accessible documents, or portions of them. This discourages information providers from publishing their content, in turn crippling the digital libraries. The idea of stealing someone's work is of course not new, but digital technology and the Internet make reproduction and distribution of documents much faster, easier and safer than the tedious paper or CD-based methods of the past.

According to the Merriam-Webster Online Dictionary [1], to "plagiarize" means

1. to steal and pass off (the ideas or words of another) as one's own
2. use (another's production) without crediting the source
3. to commit literary theft
4. to present as new and original an idea or product derived from an existing source.

1.2 Why plagiarism detection?

One salient area where plagiarism becomes a major problem is the education system. The problem of students turning to the Internet for a quick-fix homework solution which shortcuts around the time-consuming work of writing programming assignments and research papers is becoming very serious. This area has even been noted by the commercial sector already. A number of paper mill services exist, offering plagiarized papers to students, sometimes even with the added option of having the paper "customized". According to a 2002 study by McCabe [2], 10% of American college students have partially copied their assignment from the Internet without proper citation, with 5% turning in verbatim copies from web pages and term-paper mills. The situation is even worse for high school students, with the figures at 52% and 16% respectively. The numbers are rising every year, not the least because plagiarizing is becoming a common (if not acceptable) part of our educational culture. The reward for doing the hard work yourself is mostly moral, and detection and punishment of plagiators very rare.

Also notable is connection between detecting plagiarism in programming and natural languages. The latter are inherently more difficult, because a plagiarized program must, in order to retain the same semantics, also retain very similar syntax. A similar syntactical parse tree of a program is thus in itself highly indicative of plagiarism, something not true for natural languages, where the connection between syntax and semantics is much more variable and vague.

Plagiarism detection as considered in this paper is a computational means of detecting the above mentioned plagiarism violations. As such it includes copy detection, which is the most straightforward case of plagiarism that duplicates parts of documents verbatim. Copy detection is not necessarily useful strictly for unlawful violations; a possible scenario is one where user is actively sifting through documents from a particular domain. Here the 'original', or *registered* documents are simply documents that have been seen already, and the user is likely not interested in minor modifications (retransmitted or forwarded messages, different versions or editions of the same work, documents coming from mirror sites and so on). He aims to gather topically related documents, without any explicit regard to plagiarism. This is a task studied in the field of Information Retrieval (IR), and indeed in general terms plagiarism detection in digital libraries can be seen as an instance of IR.

2 Document Representation

Vector Space Model (VSM)

Information Retrieval is the part of computer science concerned with retrieving, indexing and structuring digital objects (e.g. text documents, images, videos) from collections (e.g., the Web, corpora). Although several different models have been proposed (see e.g. [3]), the one relevant for this section is the

Vector Space (VS) Model. Here objects are not represented directly, but rather approximated by *features*. What constitutes a feature is application dependent – in our case of text retrieval, most common choice are terms as delimited by white space, or term bigrams. These features are then assigned specific values for each object, leading to a representation of the object by a vector. Even though assignment of discrete values (e.g., 0, 1 or $0, 1, \dots, n$) is possible, most extensions to the basic model modify the values by weights, making the vector real-valued. Documents proximity can then be estimated by vector similarity measures, such as vector dot product, cosine similarity and others.

The VS model implies several choices: firstly, which features to extract from the documents, secondly, what weights to assign them and finally how to compute document similarity. The standard practise is to take the set of all tokens that occur in the document and note their frequencies. This tacitly assumes position independence within the document, and also independence of terms with respect to each other. This assumption is intuitively wrong (the term ‘surf’ has different meaning within the context of surfing on the Web and surfing on the beach), but empirical NLP studies nevertheless report good results using it. This approximation of a document by a set of its terms (and their frequencies) is called the *Bag of Words (BOW)* approximation.

Latent Semantic Indexing

To overcome the limitations of simple term overlap, semantic modifications of VS were introduced. One of them is Latent Semantic Indexing (LSI), a technique based on Vector Space model which aims to create associations between conceptually connected documents and terms. Research into LSI originated with [4]. LSI uses linear algebra techniques (i.e., Singular Value Decomposition, SVD), as explained in [5]. The following paragraphs give brief introduction into theoretical background and intuition into how LSI operates on textual domain. A very enticing feature of LSI is that it is a so-called unsupervised method, meaning that no explicit input of knowledge is required for training. It has been shown that LSI has good retrieval performance [10].

Let m be the rank of a term-document matrix M , which may be the TF-IDF matrix described in the previous IR section. We may decompose M into $M = U \cdot S \cdot V^T$, where U (size $t \times m$) and V^T (size $m \times d$) have orthonormal columns and S is diagonal. The columns of U (resp. V) are called the left (resp. right) *singular vectors* (or *eigenvectors*) and are the (normalized) eigenvectors of $M \cdot M^T$ (resp. $M^T \cdot M$). The values in S are the (positive) square roots of the eigenvalues of $M \cdot M^T$ (or equivalently, of $M^T \cdot M$). They are positive real numbers, because $M \cdot M^T$ is symmetric and positive definite. The decomposition process is called *Singular Value Decomposition (SVD)*¹. Without loss of generality, we can assume the positive real diagonal elements in S , called *singular values*, are sorted by their magnitude, and the corresponding *left and right eigenvectors* in U, V^T are

¹ The Singular Value Decomposition is used to solve many problems (e.g. pseudo-inverse of matrices, data compression, noise filtering) and is a least squares method. LSI uses it to find a low rank approximation of the term-document matrix M .

transposed accordingly. By keeping only the k largest singular values we can reduce S to S_k of size k . Similarly if we keep only the first k columns of U and first k rows of V^T , we get $M_k = U_k \cdot S_k \cdot V_k^T$ (with dimensionalities of $(t \cdot d) = (t \cdot k) \times (k \cdot k) \times (k \cdot d)$). This process is depicted in Figure 1. We call M_k the *rank-k approximation* of M and k the *number of factors*. In fact, as shown in [6], the Eckart-Young theorem states that M_k is the best rank- k approximation of M with respect to the Frobenius norm (2-norm for matrices). How to select the optimal number of latent dimensions k is still an open problem. However empirical studies show that values between 100 and 300 result in best text retrieval performance. In [7,8] the authors propose a statistical test for choosing the optimal number of dimensions for a given collection.

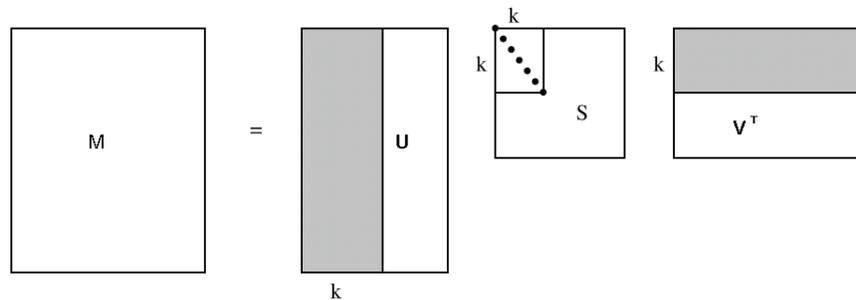


Fig. 1. LSI concept formation: Rank- k matrix approximation of M is obtained by truncating the U , S and V^T matrices from Singular Value Decomposition. Figure taken from [5].

3 DML-CZ

Czech Digital Mathematics Library (DML-CZ) [9] is a project aiming to collect historical mathematical papers and documents within the domain of Bohemia. This includes scanning and using OCR on old papers from pre-digital era. All documents are carefully processed, enriched with metadata and made accessible via web tool called Metadata editor `editor.dml.cz`. The collection, with an additional input of 15,767 articles from NUMDAM, contains 21,431 relevant articles. Out of these, there are 8,145 articles in English suitable for our experiments.

Having such collection offers interesting challenges – in what way do we let our users browse the library? All mathematical articles are reviewed, plus the group of interested people is rather narrow, so plagiarism is unlikely. But still the questions can be asked – are there any suspiciously similar documents within our library? Can document similarity facilitate and enhance browsing experience of the user?

To apply our VSM method as described above, we converted the 8,145 articles to vectors, using both TF-IDF and LSI. For LSI, we reduced dimensionality to the top 200 concepts, in accordance with common IR practice. Then we evaluated pairwise document similarity, using angle distance (cosine measure, similarity range is $(0.0, 1.0)$) and also plotted the results as 2D matrices. An example of a (part of) similarity matrix for LSI is in Figure 2.

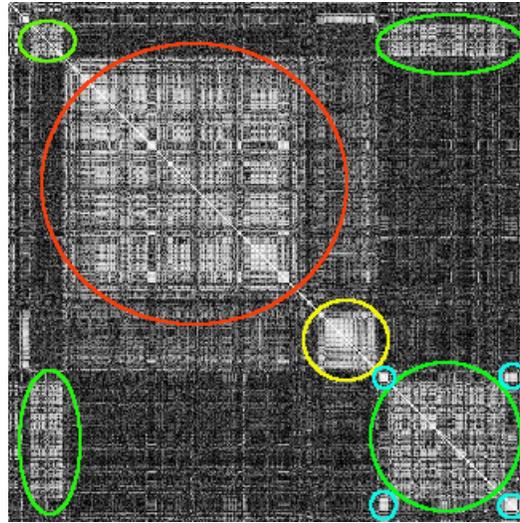


Fig. 2. An example of pair-wise document similarity on a subset of documents. Each pixel represents similarity of one pair of documents, the whiter the more similar. Note that the diagonal is necessarily white, because a document is always maximally similar to itself. The method used is Latent Semantic Indexing. See text for information on the highlighted regions.

4 Results

First question to answer is how much do TF-IDF and LSI differ on our dataset. Statistics show that the mean difference over all articles is 0.0299, with standard deviation of 0.037. Inspection reveals that in most cases the scores are indeed very similar, but there are also many pairs of documents for which the two methods vary widely, as the comparison of mean and standard deviation would suggest. See Appendix for an example of a pair of documents where TF-IDF suggested no similarity (score of 0.08) while LSI scored 0.98.

Perhaps more interesting than the numbers themselves is how well does this vector similarity translate to similarity as perceived by users. Unfortunately we do not have a referential tagged corpus of pair-wise document similarities

to compare our results against. However, thanks to the nature of our dataset, we have access to article metadata. One piece of metadata present for each of our articles is its position within MSC classification [11] hierarchy. This is a fixed taxonomy of mathematical areas to which documents are manually assigned by the author or the reviewer. In Figure 2, we selected one node in the MSC hierarchy and considered only those documents in our collection that fall into this category. The category is named *20: Group theory and generalizations* and is further subdivided into smaller categories (*20Dxx Abstract finite groups* etc.). We group documents along the axes according to these subcategories and observe how well does the suggested similarity – represented by shade of gray – correspond to subcategory clusters suggested by MSC. Although there are similarities between individual articles all over the graph, we may observe there are four main “light” clusters. These are highlighted in red, yellow, green, blue and correspond to articles from categories $20Dxx+20Exx+20Fxx$, $20.30+20Kxx$, $20.92+20.93+20Mxx$ and $20Lxx+20Nxx$, respectively. Descriptions of these ten subcategories of *Group theory and generalizations* are:

- 20.30 (1959-1972) Abelian groups
- 20.92 (1959-1972) Semigroups, general theory
- 20.93 (1959-1972) Semigroups, structure and classification
- 20Dxx Abstract finite groups
- 20Exx Structure and classification of infinite or finite groups
- 20Fxx Special aspects of infinite or finite groups
- 20Kxx Abelian groups
- 20L05 Groupoids (i.e. small categories in which all morphisms are isomorphisms). For sets with a single binary operation, see 20N02; for topological groupoids, see 22A22, 58H05.
- 20Mxx Semigroups
- 20Nxx Other generalizations of groups

Note that all of the suggested clusters are meaningful and also that the algorithm correctly linked obsolete categories 20.92 and 20.93 (used between the years of 1959 and 1972) with their new version of 20Mxx. Although these visual results cannot substitute full analytical evaluation, they are nevertheless quite encouraging.

Next step is to analyze highly similar documents for plagiarism. As mentioned above, finding actual plagiates is highly unlikely due to the nature of the domain. Indeed, analysis shows that all suspicious documents are in fact conference announcements, in memoriams and the like. If there was plagiarism present in the dataset, its complexity was beyond both LSI’s and the author’s ability to detect it.

5 Conclusion

We have presented a robust statistical method for text similarity, applied to a collection of real documents. These documents come from a digital library

of mathematical texts and also have metadata attached, which allowed us to visually compare quality of document similarity. Although our application of plagiarism detection did not yield any positive hits, it nonetheless serves as proof of concept and can be extended and used on other collections.

Acknowledgements. This study has been partially supported by the grants 1ET200190513 and 1ET100300419 of the Academy of Sciences of the Czech Republic and 2C06009 and LC536 of MŠMT ČR.

References

1. Merriam-Webster Online Dictionary, November 2008, <http://www.m-w.com/dictionary/plagiarize>.
2. McCabe, D., 2002. Cheating: Why Students Do It and How We Can Help Them Stop. *American Educator* (2001–2002) winter, pages 38–43.
3. Regis Newo Kenmogne: Understanding LSI via the Truncated Term-term Matrix. Diploma Thesis at the University of Saarland, Dept. of Computer Science, May 2005.
4. Furnas, Deerwester, Dumais, Landauer, Harshman, Streeter and Lochbaum, 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In: *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in Information Retrieval*, pages 465–480.
5. Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, pages 573–595, December 1994.
6. G. H. Golub, C. F. Van Loan, 1989. *Matrix Computations*. John Hopkins Press.
7. Zha, H., Simon, H., 1998. A subspace-based model for latent semantic indexing in information retrieval. In *Proceedings of the Thirteenth Symposium on the Interface*. pp. 315–320.
8. Ding, C. H. Q., 1999. A similarity-based probability model for latent semantic indexing. In *Proceedings of the Twentysecond Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–65.
9. Bartošek, M. and Lhoták, M. and Rákosník, J., Sojka, P. and Šárky, M., 2008. DML-CZ: The Objectives and the First Steps. In *CMDE 2006: Communicating Mathematics in the Digital Era*, AK Peters Ltd., pages 69–79.
10. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6), pages 391–407.
11. The Mathematics Subject Classification (MSC) taxonomy, November 2008, URL <http://www.ams.org/msc/>

Appendix: TF·IDF vs. LSI Differences

Below there are two articles mentioned in the text where TF·IDF and LSI scores differ dramatically.

On an unrelated note, observe the multiple OCR errors present in the text. These types of low level (character level) errors render application of more refined, semantic-based methods of text analysis very problematic. One of the

advantages of more crude, statistical methods such those based on VSM used in this paper is that these errors are not complete show-stoppers for plagiarism detection.

1. Czechoslovak Mathematical Journal, vol. 24 (99) 1974, Praha

NEWS and NOTICES IN MEMORIAM PROF. RNDr. KAREL CERNY On 15 January 1974, RNDr. Karel Cerny, Associated Professor of Mathematics at the Czech Technical University, died in Prague. Prof. Cerny was born on 6 July 1909 at Zbyslavice near Caslav. After completing his mathematical studies at Charles University in 1933 he became lecturer at the Faculty of Mechanical Engineering. He remained member of staff of the Faculty till 1953 except for the years 1942-45 when he suffered from Nazi persecution. In 1953 he was appointed Associated Professor (Dozent) first at the Faculty of Architecture and later at the Faculty of Civil Engineering of the Czech Technical University. Prof. Cerny spared no effort in his educational activity which may be characterized by his full devotion and responsible approach. The scientific interest of K. Cerny, who had been a pupil of Prof. V. Jarnik, was concentrated on the theory of numbers, particularly on the metric theory of diophantine approximations. A more detailed biography of Prof. Cerny is published in Cas. pest. mat. 99 (1974), 321 - 323. Editorial Board

2. ARCHIVUM MATHEMATICUM (BRNO) Vol. 26, No. 2-3 (1990), 65-66

THIS ISSUE OF ARCHIVUM MATHEMATICUM IS DEDICATED TO THE NONAGENERIAN OF * ACADEMICIAN OTAKAR BORtFVKA Academician Otakar Boruvka, Nestor and legend of the Brno mathematicians, long ago one of the leaders of the Czechoslovak mathematical life, a prominent representative of our science abroad, excellent teacher and oiganizer of the scientific life was ninety on May 10, 1989. In full mental freshness, creating activity, in enviable spirit, in constant interest in mathematical events. In 1920-as a student-he passed from the Czech Technical University to the newly founded Faculty of Science of the Brno University and here he passed a state examination in mathematics and physics in 1922. From the year 1921he was a lecturer in the year 1928 he became an associate professor, from the year 1934 he was a professor assistant and' from the year 1946 (with the effectivness from the year 1940) he was a regular professoi of our faculty. From the year 1970 he is a member of the Mathematical Institute of the Czechoslovak Academy of Sciences' in Brno. For the time being he is an author of 84 original scientific papers from the area of differential geometry, general algebra and differential equations and 50 further popular and bibliografical papers. For his results he was awarded a State Prize of Klement Gottwald in the year 1959 and Order of Labour in the year 1965, {hr id="0072"/> from the year 1953 he was a corresponding member and from the year 1965 a regular member of the

Czechoslovak Academy of Sciences, he is an honourable doctor of the Komensky University in Bratislava, and honourable member of the Association of the Czechoslovak Mathematicians and Physicists and he received a number of medals and diplomas of the universities and scientific associations in our country and abroad. Last but not least, he gave rise to this journal (25 years ago, in 1965) and was its first editor-in-chief. The rare life anniversary of the Academician Otakar Boruvka is of course associated with a number of summary publications in professional and popular press (e.g. Czech. Math. Journal, vol. 39 (113) 1987, 382-384). To us, belonging to the generations of his students, members of scientific seminars, founded or oriented by him, to those, inspired by his work, to his younger collaborators and colleagues and to those esteeming his character, is, however, this reality not only a reason for valorizing his admirable work but also for an opportunity to express our homage to our honoured person by the results of our works. We wish to Academician Boruvka health and good humour in order to be able to give away, in further years, from the treasury of his wisdom and experience. Photo: J. France