Jan PLATOS, Michal PRILEPOK, Vaclav SNASEL

VSB-Technical University of Ostrava

Text comparison using data compression

***Abstract**. Similarity detection is very important in the field of spam detection, plagiarism detection or topic detection. The main algorithm for comparison of text document is based on the Kolmogorov Complexity, which is one of the perfect measures for computation of the similarity of two strings in defined alphabet. Unfortunately, this measure is incomputable and we must define several approximations which are not metric at all, but in some circumstances are close to this behaviour and may be used in practice.*

***Streszczenie.** W artykule omówiono metody rozpoznawania podobieństwa tekstu. Głównie używanym algorytmem jest Kolmogotov Complexity. Głównym ograniczeniem jest brak możliwości dane algorytmu są trudne do dalszego przetwarzania numerycznego – zaproponowano szereg aproksymacji. (**Porównanie tekstu przy użyciu kompresji danych**)*

Keywords: data compression, text similarity, normalized compression distance, Kolmogorov complexity.
Słowa kluczowe: kompresja danych, porównanie tekstu..

Introduction

The growing number of documents, tests, books and scientific papers brings new challenges in the area of content mining, text processing and understanding and author identification or confirmation. One of the interesting tasks is also plagiarism detection. This problems is actual in many areas such as patent applications, program's source codes copying, image usage without permission, DNA processing, and many others.

This article is focusing of the overview of algorithms for the comparison of text documents using data compression. This task is investigated very long time but it becomes even more acute with the massive expansion of the personal computers in the world. The comparison of text documents is highly related to the term plagiarism. The plagiarism may be defined using several definition but we are following this one: The plagiarism detection is the identification of highly similar sections in texts or other objects [1]. Other definitions may be found in the literature such as this [2]. The plagiarism detection may be divided into two major areas - external and intrinsic [1]. The External plagiarism is defined as an identification of the part of the document *d* which exists in any of the document in the document collection *D*. The Intrinsic plagiarism detection is a method the possibly plagiarized pars of the documents just from the document itself. The second one is more complicated. Other methods of text comparison which are not covered in plagiarism detection may be described as follows; comparison of the meaning of the texts in the same or different languages, categorization of the texts, etc.

Text comparison methods

The main problem of the comparison of text documents is the definition of the similarity or dissimilarity measure. The most suitable methods for similarity measure are metrics [4]. The distance is formally defined as a function over Cartesian product over set X with non-negative real value [5] and [6]. The metric is a distance which satisfies three conditions for all $x, y, z \in X$:

1. D(x, y) = 0 if and only if x = y
2. D(x, y) = D(y, x)
3. D(x, y) ≤ D(x, y) + D(y, z)

The condition 1 is called identity, condition 2 is called symmetry and condi- tion three is the triangle inequality. This definition is valid for any metric, e.g. Euclidean Distance, but the application of this principle into document or data similarity is much complicated.

The basic ideas were suggested and defined in related works by Li et al. [6], and Cilibrasi and Vitanyi [5]. They defined the so-called Normalized Information Distance (NID). The NID is based on the definition of the Kolmogorov complexity (KC): The Kolmogorov complexity K(x) of the string x = {0, 1}* is the length of the shortest binary program with no input that outputs x [5]. The Kolmogorov complexity of the two strings may be expressed as follows: The Kolmogorov complexity of x given y is the length of the shortest binary program, for the reference universal prefix Turing machine, that on input y outputs x; it is denoted as K(x|y) [5]. The NID is then defined as follows:

$$NID(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\min\{K(x|y), K(y|x)\}}$$

Unfortunately, the Kolmogorov complexity function is non-computable. But Li et al. and Cilibrasi reformulated this problem into a computable form using the replacement of the Kolmogorov complexity by using data compression [5,6]. The non-metric measure developed from their work is a Normalized Compression Distance.

Normalized Compression Distance

The Normalized Compression Distance (NCD) is based on Kolmogorov complexity. It makes use of standard compressors in order to approximate Kolmogorov complexity. Several papers have already used this similarity in order to compare texts of different kinds and in different ways. The NCD has been used for text retrieval [7], text clustering, plagiarism detection [8], music clustering [9], music style modelling [10], automatic construction of the phylogeny tree based on whole mitochondrial genomes [11], the automatic construction of language trees [12, 6], and the automatic evaluation of machine translations[13].

The NCD is a mathematical way for measuring the similarity of objects. Measuring of similarity is realized by the help of compression where repeating parts are suppressed by compression. NCD may be used for comparison of different objects, such as images, music, texts or gene sequences. NCD has requirements to compressor. The compressor meets the condition C(x) = C(xx) within logarithmic bounds [14]. We may use NCD for detection of plagiarism and visual data extraction [15, 5].

The resulting rate of probability distance is calculated by the following formula:

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

where

- C(x) ist he length of compression of x.
- C(xy) is the length of compression concatenation of x and y.

- min{x, y} is the minimum of values x and y.
- max{x, y} is the maximum of values x and y.

The NCD value is in the interval 0 ≤ NCD(x, y) ≤ 1 + E. If NCD(x, y) = 0, then files x and y are equal. They have the highest difference when the result value of NCD(x, y) = 1 + E. The constant E describes the inefficiency of the used compressor. The NCD is not a metric. It is an approximation of the NID. The computation of the NCD is very efficient because we do not need to create the output itself. We compute only the size of the output. A study of the efficient implementation of the compression algorithms may be found in [16].

Pattern Representation Scheme using Data Compression

A different approach which uses a Data Compression for Similarity detection was suggested by Watanabe in [24].The approach is called Pattern Representation Scheme using Data Compression, or PRDC. The input data are converted into a string representation. This string representation is then compressed by the set of encoding dictionaries. Each encoding dictionary produces one value of Compressed Ration Vector. This vector is then used by the standard Vector Quantization technique as a feature vector. The Authors presented several experiments which confirm the ability of PRDC do detect patterns (similar or same parts) in input data. First experiment consists in categorization of Music and Voice in audio files. These files are pre-processed by the segmentation of the records into short frame of constant size. The results show that the algorithm was able to categorize the samples into hierarchical structure. The Second experiment deals with prediction of the function and structure from the DNA or Amino Acid sequence. 33 tested sequences were classified into three clear groups as a result. The last experiments deal with processing of images. First sketches were classified according to simple lines. The PRDC was able to retrieve sketches according the example query with 90% accuracy. The second experiment classifies colour satellite images according to type of land-cover. The PRDC is very fast but is miss the usage of the join step which is included in NCD measure.

Fast Compression Distance

The Fast Compression Distance (FCD) combines the speed of the PRDC and the join factor of the NCD together [25]. The approach is based on the Dictionary based compression LZW which is a version of the Lempel-Ziv algorithm suggested in 1978. This algorithm is able to extract a dictionary of phrases from the string input. Each compared object is converted into string representation at the beginning. Then, a dictionary is extracted by the LZW algorithm and lexicographically sorted. The sorting enables fast set operation such a union and intersection as well as binary search operation. Then we may compute the FCD by the following equation:

$$FCD(x,y) = \frac{|D(x)| - |\cap(D(x), D(y))|}{|D(y)|}$$

The D(x) and D(y) means a dictionary extracted from object x and y. The |X| is the size of the set X and ∩ is and intersection. The FCD(x,y) ranges from 0 to 1, where 0 means minimum distance and 1 is the maximum distance. The intersection between dictionaries represents the join step which is also used in NCD. The FCD is much faster than the NCD because the NCD needs the compression of the joined file and the FCD need just to compute the intersection of the dictionaries. The dictionaries itself may be extracted only once when a new object is found and ready to compare. The authors test the FCD measure on the CBIR (Content Based Information Retrieval) software on images. The FCS was more than 14 times faster than NCD and achieved accuracy better than 97% with accuracy 77% of NCD.

Compression Based Dissimilarity

A similar approach to NCD was published by the Keogh [27] and is called Compression Based Dissimilarity (CDM). The CDM is defined by the following equation.

$$CDM(x,y) = \frac{C(xy)}{C(x) + C(y)}$$

As may be seen, the equation is defined in similar way as the NCD but the maximum and minim is replaced the simple sum. The results achieved by the CDM are very similar to the NCD. Authors performed set of experiments to proof the efficiency of the CDM for several tasks. One task was clustering of the EEG signals, hierarchical clustering of the different species according to their DNA and outlier detection on the translation of the bible into several different languages. Results of all experiments show that CDM is able to solve the defined task very efficiently.

Application of the Compression on Text Similarity

Despite the application mentioned with the defined measures, other applications of the similarity detection were also published.

Ferragina et al. [28] published a paper where biological sequences and structures were classified using NCD measure. The PPM algorithm as well as GZIP was used as compressors for the different type of data. But they used also tents of algorithm and variants on different types of data. The achieved results were nice and algorithm shows that the NCD is adequate for analysis of biological data mainly because of its flexibility and scalability with data set size.

A nice study which compares many different compressors on the image data using NCD was published by the Vázquez and Marco [29]. They compare image compressors and universal compressors on several images on various data formats. The image compressors JPEG and JPEG 2000 were proved as useful when comparison of two images is needed. The reason is that these algorithms are not able to use any information from one image to compress another more efficiently. In the contrary, the GZIP and PKZip algorithms was able to detect similar images very efficiently, except the images stored in JPEG and JPEG 2000 file format, because these file formats compress the images very well and the compression achieved by the compressors is almost zero. The similar results were achieved with the block oriented algorithm as well as with the context based algorithms.

Very similar experiments were performed by the Pinho and Ferreira [30]. They test many compression algorithms on the grayscale images. The achieved results were very similar to the previous one. Three different image compressors have very poor results and were not able to identify any similarity even on the same images. The other algorithms which were based on the dictionary and context were better, but the results of more powerful algorithms LZMA and PPMD were worse than the GZIP when the similarity of two images is compared using NCD.

Conclusion

This paper presents and overview of the many different measures which may be used for detection of the similarity between text documents. Any of the measure may be used in combination with the proper compression algorithm for detection of the patterns, plagiarism detection and/or computation of the similarity of two documents. Moreover, these methods may be used for comparison of multimedia when they are converted into text form such as images, audio files or biological structures and sequences.

This work was partially supported by the Grant Agency of the Czech Republic under grant no. P202/11/P142, and by the Grant of SGS No. SP2013/70, VSB - Technical University of Ostrava, Czech Republic, and was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

REFERENCES

[1] M. Potthast, B. Stein, A. Eiselt, B. universitt Weimar, A. Barrn-cedeo, and P. Rosso, "P.: Overview of the 1st international competition on plagiarism detection," in In: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org, 2009, pp. 1–9.

[2] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - a survey."

[3] J. Platos, V. Snasel, and E. El-Qawasmeh, "Compression of small text files," Ad- vanced Engineering Informatics, vol. 22, no. 3, pp. 410–417, 2008.

[4] A. Tversky, "Features of similarity," Psychological Review, vol. 84, no. 4, pp. 327– 352, 1977, cited By (since 1996)1968.

[5] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by compression," IEEE Transactions on Information Theory, vol. 51, no. 4, pp. 1523–1545, 2005.

[6] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," IEEE Transactions on Information Theory, vol. 50, no. 12, pp. 3250–3264, 2004.

[7] A. Granados, "Analysis and study on text representation to improve the accuracy of the normalized compression distance," AI Commun., vol. 25, no. 4, pp. 381–384, 2012.

[8] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, "Shared information and program plagiarism detection," IEEE Transactions on Information Theory, vol. 50, no. 7, pp. 1545–1551, 2004.

[9] R. Cilibrasi, P. Vitanyi, and R. de Wolf, "Algorithmic clustering of music based on string compression," Computer Music Journal, vol. 28, no. 4, pp. 49–67, 2004, cited By (since 1996)76.

[10] S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano, "Using machine-learning methods for musical style modeling," Computer, vol. 36, no. 10, pp. 73–80, 2003, cited By (since 1996)25.

[11] M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information- based sequence distance and its application to whole mitochondrial genome phy- logeny," Bioinformatics, vol. 17, no. 2, pp. 149–154, 2001, cited By (since 1996)274.

[12] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," Physical Review Letters, vol. 88, no. 4, pp. 487 021–487 024, 2002, cited By (since 1996)145.

[13] J. J. Vayrynen, T. Tapiovaara, K. Kettunen, and M. Dobrinkat, "Normalized com- pression distance as an automatic MT evaluation metric," in Proceedings of MT 25 years on, 21–22 Nov 2009 Cranfield, UK, to appear.

[14] D. Sculley and C. Brodley, "Compression and machine learning: A new perspective on feature space vectors," 2006, pp. 332–341, cited By (since 1996)17.

[15] P. M. B. Vit´anyi, "Universal similarity," CoRR, vol. abs/cs/0504089, 2005.

[16] J. Walder, M. Kratky, R. Baca, J. Platos, and V. Snasel, "Fast decoding algorithms for variable-lengths codes," Inf. Sci., vol. 183, no. 1, pp. 66–91, 2012.

[17] D. Kirovski and Z. Landau, "Generalized lempel-ziv compression for audio," in Multimedia Signal Processing, 2004 IEEE 6th Workshop on, 2004, pp. 127–130.

[18] V. Crnojevic, V. Senk, and Z. Trpovski, "Lossy lempel-ziv algorithm for image compression," in Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2003. TELSIKS 2003. 6th International Conference on, vol. 2, 2003, pp.522–525 vol.2.

[19] D. Chuda and M. Uhlık, "The plagiarism detection by compression method," in CompSysTech, B. Rachev and A. Smrikarov, Eds. ACM, 2011, pp. 429–434.

[20] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate cod- ing," Information Theory, IEEE Transactions on, vol. 24, no. 5, pp. 530–536, 1978.

[21] M. Potthast, B. Stein, A. Barr´on-Ceden˜o, and P. Rosso, "An Evaluation Frame- work for Plagiarism Detection," in 23rd International Conference on Computa- tional Linguistics (COLING 10), C.-R. Huang and D. Jurafsky, Eds. Stroudsburg, Pennsylvania: Association for Computational Linguistics, Aug. 2010, pp. 997–1005.

[22] J. Sammon, "A nonlinear mapping for data structure analysis," Computers, IEEE Transactions on, vol. C-18, no. 5, pp. 401–409, 1969.

[23] R. Arnold and T. Bell, "A corpus for the evaluation of lossless compression algo- rithms," pp. 201–210.

[24] Watanabe, T.; Sugawara, K.; Sugihara, H., "A new pattern representation scheme using data compression," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.5, pp.579,590, May 2002

[25] Daniele Cerra, Mihai Datcu, A fast compression-based similarity measure with applications to content-based image retrieval, Journal of Visual Communication and Image Representation, Volume 23, Issue 2, February 2012, pp 293-302, ISSN 1047-3203

[26] T.A. Welch, A technique for high-performance data compression, Computer 17 (6) (1984) 8–19.

[27] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Towards parameter-free data mining. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 206-215. 2004.

[28] Ferragina P et al., Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment, BMC Bioinformatics 2007, 8:252

[29] Pere-Pau Vázquez and Jordi Marco, Using Normalized Compression Distance for image similarity measurement: an experimental study, The Visual Computer, November 2012, Volume 28, Issue 11, pp 1063-1084

[30] Pinho, A.J.; Ferreira, P.J.S.G., "Image similarity using the normalized compression distance based on finite context models," Image Processing (ICIP), 2011 18th IEEE International Conference on , vol., no., pp.1993,1996, 11-14 Sept. 2011

***Authors**: Jan Platos, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:jan.platos@vsb.cz; Michal Prilepok,VSB-Technical University of Ostrava, 17 listopadu 15,70833 Ostrava Poruba, E-mail:michal.prilepok@vsb.cz; Vaclav Snasel, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:vaclav.snasel@vsb.cz;.*