INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGIES, ENGINEERING AND MANAGEMENT SCIENCE

Document similarity measurement techniques overview

Vladimír Hanušniak
Faculty of Management Science and Informatics, University of Zilina, Slovak Republic
vladimir.hanusniak@uvp.uniza.sk

Abstract

Examine data for similar items is one of the fundamental data-mining problem. Application of methods for similarity search could be useful for plagiarism or near-duplicate web page detection. The basis to measure similarity are distance measurement methods. The paper describes several distance measure methods and document representation suitable for document similarity analysis. Conclusion of the paper state the recommendation to measure document similarity.

Keywords: document similarity, data mining, plagiarism detection, document analysis, string

Introduction

Over the past few years, Internet became the biggest publicly available information storage. Digital era caused that many publicly available documents in digital form are easily available and readily reproducible. Organizations store huge collections of documents and emails with multiple copies and versions that may not be kept up to date. It is often essential to determine which one is the current copy and which one is the draft. However, even bigger problem is copying another's work or borrowing someone else's original ideas also known as plagiarism. More serious is occurrence of this problem in research and scientific papers. Increasing interest in the problem is confirmed by the number of publications on this matter in recent years. This paper describe well known fundamental methods used for measuring similarity of document. It give us an overview for analyzing and adjusting the methods parameters in order to achieve higher accuracy.

Plagiarism taxonomy

There are no two humans, no matter what languages they use and how similar thoughts they have, write exactly the same text. Thus written text, which is stemmed from different authors should be unique to some extent, except for cited portions [3]. Quoting is required for any borrowed content and original author should be listed in the reference list.

Plagiarism taxonomy and patterns are described in many research papers and a number of publication concerning plagiarism in research papers were published in recent years [4, 5, 6]. Most of the proposed techniques rely on substring matching, keyword similarity or fingerprint analysis [7]. Basic plagiarism detection taxonomy is depicted on Figure 1.

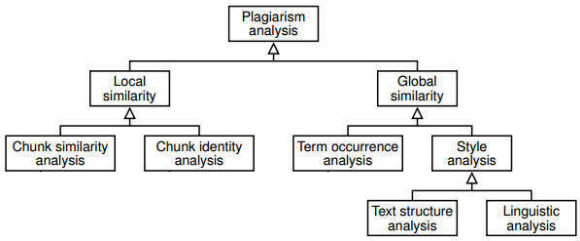

Figure 1 - Plagiarism detection techniques

This paper is focused on substring matching technique. The most important part of the technique is to measure similarity between two objects. There are several distance measure methods and the most famous are describe in the following section.

Distance measures

Distance function, also called metric is function that define distance between each pair of elements of given set. Suppose we have a set of points X, called a space. A distance measure on this set is a function $d(x,y)$ that produces a non-negative real number and for all x, y, z in X, the following conditions are satisfy:

1. $d(x, y) \geq 0$ (*no negative distances*).
2. $d(x, y) = 0$ if and only if $x = y$ (coincidence

axiom).
3. $d(x, y) = d(y, x)$ (*distance is symetric*).
4. $d(x, z) \leq d(x, y) + d(y, z)$ (*the triangle inequality*).

Triangle inequality is the most complex and it says, that we cannot obtain any benefit if we are forced to travel from A to B via some particular third point C [1].

Euclidean distance

One of the most familiar distance measure is Euclidean distance. It is one we normally think of as "distance" and define distance of two objects in Euclidean space. The distance *d(p,q)* is given by the formula:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}. \quad (1)$$

It squares the distance in each dimension, sum the squares and takes the positive square root.

Manhattan distance

The simple sum of the horizontal and vertical components called Manhattan distance. It is used in Euclidean space. Euclidean distance in general is given by formula:

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{1/r} \quad (2)$$

Euclidean distance is defined for the constant r=2, however *Manhattan distance* is defined for the constant r=1. The name of this distance come from well-known gridlike street geography of the New York borough of Manhattan.

Jaccard distance

The most widely used distance in data-mining is called *Jaccard distance*. Similarity measure of two sets is done by looking at the relative size of their intersection (union). If we define two sets A and B, the *Jaccard similarity* is given by formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

Jaccard distance is complementary to *Jaccard similarity* is obtained by subtracting the *Jaccard similarity* from 1. It is the same as dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (4)$$

Example:
D_1 = {"I", "went","to","school", "today"}
D_2 = {"I", "went","to","work", "yesterday"}

$J(D_1,D_2)$ = 3 / (10-3) = 3 / 7 = **0.4285**

Cosine distance

The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (*time frequency*) of each document, but the angle between the documents [2]. Figure 2 shows vector space model of documents modeled as a vectors.

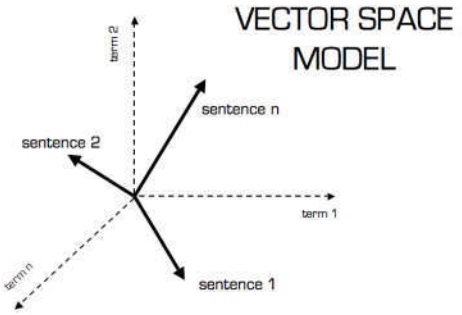

Figure 2 - Vector space model [2]

Cosine of an angle between two vectors x and y is the dot product *x.y* divided by their *Euclidean distance* (see formula x).

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (5)$$

The angle of two vectors which pointing to two far distance points could be also small. This characteristic is important while using Cosine distance. Suppose we have two documents with the same word "car". First of them contains this word 10

times, the second one 30 times. Hence, Euclidean distance between these documents is high, but the angle between document vectors is small because they are pointing to the same direction.

Edit/Leveshtein distance

This distance makes only sense when documents are strings. It is also used in DNA sequence analysis. The distance between two strings $x = x_1x_2 \ldots x_n$ *and* $y = y_1y_2 \ldots y_n$ is the smallest number of insertion and deletion of single characters that will convert x to y [1]. In addition, Leveshtein distance permit character substitution.

An example below shows basic Edit and Levenshtein distance measure between two words – "*kitten*" and "*sitting*". Edit distance of these words is 5 (two deletion and three insertion), but Levenshtein distance is 3 (two substitution and one insertion).

Distance measure conclusion

A couple of distance measurement metrics was described. Each one of them is useful for different type of analysis.

Document representation

To find out similarity, the documents should be represented in the form which ensure time efficiency processing. The fundamental method is to compare the documents character by character. However, it is time consuming and it can find only exact copies. If the documents are different only in one character, it won't mark them as similar. However, similar documents could differ in few words but also in half of the context. Hence, the better approach is to represent the documents as a set of short strings that appear within it.

There are two approach to create these short string also called shingles - as a set of words or a set of characters. Identifying of lexically similar documents could not be precise enough while using set of words. The most effective way is to represent document as a set of short strings that appear within it [1]. If we do so, documents that share pieces as short as sentences or even phrases will have many common elements in their sets, even if those sentences appear in different orders in the two documents [1]. Another approach is to represent document as a set of consecutive words with given length.

Let the document to be represented as a set of five consecutive words:

$$W = \{w1, w2, w3, w4, w5\} \quad (6)$$

Then we can represent the document as set of 3-shingles:

$$S = \{\{w1,w2, w3\},\{w2,w3, w4\},\{w3,w4, w5\}\} \quad (7)$$

The document representation could contain only the most frequent words within document as *tf-idf* does [1]. Then it is represented as n-dimensional vector wt. Let $wf(d, w)$ denote the frequency of word w in document d:

$$wd = (wf(d, w_1), wf(d, w_2), wf(d, w_3), \ldots, wf(d, w_n)) \quad (8)$$

The rule used in *tf-idf* is – "The more frequent word within the document, the more important it is."

Many preprocessing operation could be applied to document before analysis:

- lowercase,
- remove punctuation,
- remove stop words e.g.: "the", "a", "to", ….,
- etc.

Storage optimization:

Storing and comparing many documents as a strings is computationally expensive. Better approach is to break down the document to a set of shingles and store them as an integer values instead of strings. Hash function is designed for this purpose. MinHash and Local Similarity Hashing technique are desire to tackle effective document similarity measures [1].

Conclusion

This paper provides a short overview of known methods to measures document similarity. Each of them is relevant to achieve different goals. If the number of single word frequency in the document is irrelevant, Cosine distance is useful to measure similarity. Computational performance and accuracy could be increased by appropriate document representation. To deal with huge amounts of documents, it could be useful to use MinHash or Local Sensitive Hashing and distance measure methods then used for results verification. The computerized methods developed during last years have mainly focused on English language. Hence, our future research will be focused on practical comparison of described measurement techniques on Slovak language. It contains many language dependent exceptions which create a space for tuning to achieve higher accuracy of the methods.

Acknowledgements

This paper is supported by the following project: University Science Park of the University of Zilina (ITMS: 26220220184) supported by the Research&Development Operational Program funded by the European Regional Development Fund.

References

[1] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press; 2 edition (December 29, 2014).

[2] Machine Learning: Cosine Similarity for Vector Space Models, [Online]. Available: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/. [Accessed 22.6.2016]

[3] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods," in IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. vol. PP, 2011, pp. 1-17

[4] Stein, B. and Eissen, S.M. (2006) Near Similarity Search and Plagiarism Analysis. In: Spiliopoulou, et al., Eds., From Data and Information Analysis to Knowledge Engineering Selected Papers from the 29th Annual Conference of the German Classification Society (GfKl) Magdeburg, Springer, Berlin Heidelberg, 430-437.

[5] Ceska, Z., M. Toman, and K. Jezek. 2008. "Multilingual Plagarism Detection." In *Artificial Intelligence: Methodology, Systems, and Applications. 13th International Conference, AIMSA 2008, Varna, Bulgaria, September 2008, Proceedings*, edited by R. Dochev, M. Pistore and P. Traverso, 83–92. Heidelberg: Springer.

[6] A. Huang, Similarity measures for text document clustering, in: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.

[7] Dan Gusfield, Algorithms on strings, trees, and sequences: computer science and computational biology, Cambridge University Press, New York, NY, 1997

[8] C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," in Proc. SEPLN, Donostia, Spain, 2012, pp. 10–18.

[9] MinHash for dummies, Available: http://matthewcasperson.blogspot.sk/2013/11/minhash-for-dummies.html [Online], [Accessed 27 10 2014].