Application of statistical N-gram clustering algorithm on words in the Macedonian language

Bojan Ilijoski
Faculty of Computer Science and Engineering
Skopje, Macedonia

Zaneta Popeska
Faculty of Computer Science and Engineering
Skopje, Macedonia

Abstract

The N-gram algorithm, one of the most famous algorithms used for statistical clustering of words, determines the similarity between two words, based upon a statistical analysis. In this paper we present the work of this algorithm and the results obtained on clustering 10 000 words in the Macedonian language. The realization of this algorithm is made in the programming language Java.

I. Introduction

One of the most important elements of the process of text mining is the pre-processing. It implies division of the text into tokens, their analysis and processing. This suggests that words which are not significant for the text will be thrown out (e.g. prepositions, conjunctions, exclamations etc.). Also, the similarity in the meaning of words, the word class a word belongs to, singularity, plurality, the tense (if it is a verb) etc., are part of the pre-processing. Another very important step is stemming. Stemming is a process through which one can find the word's stem. Word stem is the part of a word which presents the base upon which different word forms can be created. In a way, it is a start point for the process of word formation. This enables clustering of morphologically similar words. Till now, many ways of finding the stem of a word have been proposed. These methods are grouped according to the way in which they function. Their more general division is into two methods, manual and automatic. In our interest are the automatic methods. Furthermore, the automatic methods are divided into truncating i.e. affix removal, statistical and mixed [1][2]. Representatives of the truncating methods are Lovins' and Porter's. These algorithms are based on rules of affixes removal. At first, these algorithms were developed for English, and then for other languages in which these sort of rules can be defined. It is really complicated to find these rules in the Macedonian language, because the result may be very complex and may include a huge number of them. Unlike truncating, the statistical methods can be applied to many languages because they are much less dependent on the morphological structure of the words in a language. As the name itself implies, these methods are based on a statistic about how often a token, a part of a word, appears in a particular word. In this way, they make the comparison between words, how big is the similarity between words. Representatives of the group are: N-gram, HMM (Hidden Markov Model) and YASS (Yet Another Suffix Striper). The representatives of the mixed methods include both approaches (truncating and statistical). Until now, N-gram algorithm has been implemented in many world languages, but as far as we know it has not been implemented in Macedonian language. The words in Macedonian are formed in such a manner that the use of this algorithm to mutually compare them, promises good results.

II. The algorithm

Adamon and Borehem developed an algorithm for finding similarity between pair of two documents [3]. The N-Gram represents the division of a word into character sequences with length n. The similarity between two words is usually computed by finding the ratio between similar n-grams in the words and sum of n-grams of the words. Van Rijsbergen proposed several ways of computing similarity between two words such as cosine, dice, jaccard, overlap and simple[4].

N-gram length		
2	3	4
_к	__к	___к
кн	_кн	__кн
ни	кни	_кни
иг	ниг	книг
га	ига	нига
а_	га_	ига_
	а__	га__
		а____

Table 1: N-gram of the word книга (_means blank space)

n	word	n-grams	dice similarity
2	книга	_к, кн, ни, иг, га, а_	0.55
	кога	_к, ко, ог, га, а_	
3	книга	__к, _кн, кни, ниг, га_, а__	0.75
	книги	__к, _кн, кни, ниг, ги_, и__	

Table 2: Two words similarity by N-gram algorithm

* Acknowledgement: The work has been supported by the FCSE.

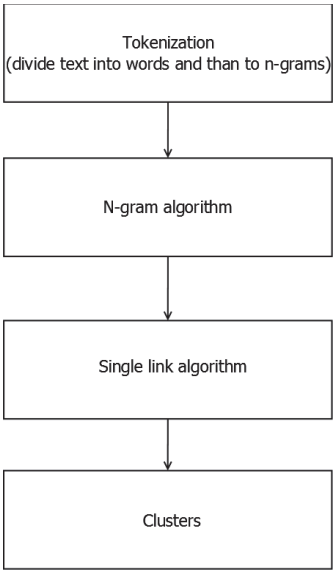

Figure 1: An algorithm work flow

A variation of an n-gram length will produce different similarities between words. Lower n sometimes will produce huge similarity between different words, only because the n-grams in one word are permutations of the n-grams in another word. Higher n will produce no similarity between similar words whose length is significantly different. This is one of the disadvantages this technique has because regardless of n's height, it can give incorrect results. Optimal value for n does not exist; usually the best chosen n depends on the word structure in languages. After the process of finding similarity between all words in a text is completed, we create a similarity matrix. Similarity matrix is a lower triangular matrix filled with similarity values between words. The next step is finding the similar words, which means grouping the words in clusters. A single link algorithm is used for this purpose[4]. This algorithm creates a graph. Every word in the graph is a node. A single node in the graph can be linked only with the nod that is most similar to it. There will be an edge between some of the nodes only if two words' similarity is bigger than some specified threshold. Every set of linked nodes in a graph creates a cluster. The fact that if in a connected sub graph (cluster) there is a path with long distance, there is a possibility that the first and the last word in this path will be very different, is a big disadvantage of this algorithm. This figure 2 represents the graph gotten by single link algorithm applied on the words став, станав, состав, состави, составни, страв, спав with threshold greater than 0.3. As you can see all the words are grouped in the same cluster because the difference between two of them is greater then the threshold. But, the similarity between спав and сосватни is only 0.07. This two words are very different but this algorithm can't see that because it made a link only to one node (word) of the cluster. That one which is the most similar to the word which we are trying to cluster.

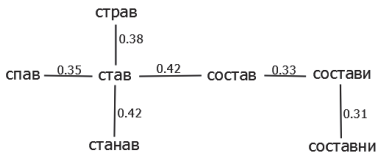

Figure 2: Single Link Algorithm Graph

This can be avoided by computing the average distance from one nod to all other members of the group, and then comparing it to the threshold. Another possibility to avoid it is by running again a single link algorithm in the cluster with higher threshold. Nevertheless, this algorithm works well when a good threshold is chosen.

III. Results

This algorithm has been tested on a set of around 10 000 words from Macedonian with different parameters of n-gram length and threshold for word similarity. Around 2600 of the words were distinct and the rest were their repetitions. The results are verified manually.

n-gram size	similarity type	threshold	no. of clusters	average size per cluster	max words per cluster	min words per cluster
2	COSINE	0.15	84	30.9	499	1
2	COSINE	0.25	618	4.2	144	1
2	COSINE	0.35	1541	1.7	44	1
2	DICE	0.15	17	152.8	945	1
2	DICE	0.25	56	46.4	425	1
2	DICE	0.35	125	20.8	247	1
2	JACCARD	0.15	57	45.6	425	1
2	JACCARD	0.25	268	9.7	133	1
2	JACCARD	0.35	667	3.9	56	1
2	OVERLAP	0.15	10	259.8	1673	1
2	OVERLAP	0.25	43	60.4	830	1
2	OVERLAP	0.35	70	37.1	624	1
2	SIMPLE	1	32	81.9	1026	1
2	SIMPLE	2	147	17.7	480	1
2	SIMPLE	3	490	5.3	379	1
3	COSINE	0.15	313	8.3	172	1
3	COSINE	0.25	1338	1.9	35	1
3	COSINE	0.35	2394	1.1	5	1
3	DICE	0.15	37	70.2	711	1
3	DICE	0.25	120	21.7	350	1
3	DICE	0.35	349	7.4	170	1
3	JACCARD	0.15	129	20.1	339	1
3	JACCARD	0.25	600	4.3	78	1
3	JACCARD	0.35	1167	2.2	35	1
3	OVERLAP	0.15	20	129.9	1327	1
3	OVERLAP	0.25	82	31.7	613	1
3	OVERLAP	0.35	212	12.3	229	1
3	SIMPLE	1	54	48.1	665	1
3	SIMPLE	2	222	11.7	322	1
3	SIMPLE	3	653	4.0	116	1

Table 3: algorithm results

* Acknowledgement: The work has been supported by the FCSE.

	N-gram length			
Threshold	DICE	COSINE	JACCARD	OVERLAP
$0.0 \leq t < 0.1$	2769895	3221177	3170121	2653770
$0.1 \leq t < 0.2$	425714	91912	128057	474537
$0.2 \leq t < 0.3$	95510	8637	19349	140107
$0.3 \leq t < 0.4$	19970	1177	3012	30553
$0.4 \leq t < 0.5$	7261	135	1366	14044
$0.5 \leq t < 0.6$	2951	4	930	7049
$0.6 \leq t < 0.7$	1191	0	172	1672
$0.7 \leq t < 0.8$	463	0	29	833
$0.8 \leq t < 0.9$	85	0	5	428
$0.9 \leq t < 1.0$	2	0	1	35
$t = 1.0$	2579	2579	2579	2593

Table 4: distribution of similar words

A. Example 1

The following clusters are produced by the algorithm with the following parameters: n-gram size: 2, similarity type: COSINE, threshold: 0.45

таква, така, тука, ваквата, таа, вака, ваква

вас, нас, јас, час

општи, оти, очи, оче, чии

B. Example 2

The following clusters are produced by the algorithm with the following parameters: n-gram size: 3, similarity type: DICE, threshold: 0.45

таква, такви, така, тука, какви, таквите, таа, качи

граматики, граматика, јазикграматики, граматичка, граматички, граматичко, граматичките, граматиката

кодификацијата, деклинацијата, деклинација, конструкцијата, конструкциите, класификикација, класифицираат, конструкција, класификации, модификации, конструкции, констатација, класификација, класификацијата, конотација

References

[1] Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms," IEEE Transactions on Aerospace and Electronic Systems,Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938

[2] Ilia Smirnov, "Overview of Stemming Algorithms," DePaul University, 2008

[3] Adamson, G.W. and Boreham, J. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles

[4] C. J. van RIJSBERGEN, INFORMATION RETRIEVAL

* Acknowledgement: The work has been supported by the FCSE.