Analysis of relation types in semantic network used for text classification

Victor Potaraev
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
vic229@rambler.ru

***Abstract*—Contemporary information systems contain lots of textual information. One of important kinds of textual information processing is text classification. Semantic network is a model which can be used to resolve different tasks including text classification. There are other models which can resolve the same task, and some of them show relatively good results, but using semantic networks has such advantages as human readability and analyzing actual semantic relations between words. Semantic networks have different set of relation types depending on the network purpose. This article is devoted to looking for a set of relation types to be leveraged in semantic network created for text classification. The analysis is performed by generating semantic networks for Russian-language texts leveraging different sets of relation types. Generated networks are used for text classification. Texts were taken from books on several technical disciplines. Proposed algorithms can be used to perform text classification when performing such tasks as dividing electronic messages on categories, spam filtering, text topic recognition and other.**

***Keywords*—semantic network, text classification, text categorization, natural language processing, semantic analysis, machine learning, data analysis.**

I. Introduction

During several previous decades humankind created a large amount of text documents. As a result, it is very important to have ability to perform automatic text classification and natural language processing. Detailed research in the area of automated text classification has started quite long time ago. One of the first researches was performed in 1961 [1] and was based on statistical method of documents indexing.

Primary purpose of text classification is to divide an unstructured set of documents into groups according to their content. Text classification can be used in the following areas:

1) Divide electronic messages on categories.
2) Spam filtering.
3) Text topic recognition.
4) Other.

There are two primary approaches to text classification and topic analysis: frequency analysis and semantic analysis. The first one is based on calculating frequency of words in text and the second is based on meaning of words (more precise). There are also methods mixing characteristics of these two approaches, such as frequency and context analysis [2].

Semantic network is an oriented graph which reflects concepts and relations between them [3]. Semantic network describes knowledge using networking structure [4]. For example, in case of 3 concepts related to each other the network will look like shown on Fig. 1.

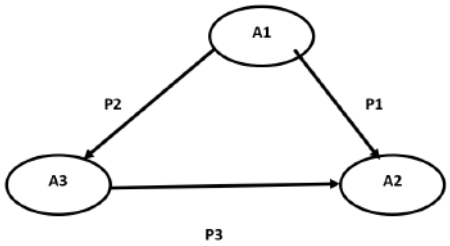

Figure 1. Example of semantic network.

Each relation in semantic network can have a kind, and kinds of relations used in network are usually selected based on specific problem being resolved [5].

There is a lot of different approaches to text classification. One of them is Naive Bayes classifier which shows good results even comparing to more complicated approaches. Other popular approaches are neural networks, support vector machines, regression methods and other [6]. Many of used models are based on working with numbers rather than actual semantic relations between words. Semantic network is a model which has such advantages as human readability and reflecting actual semantic relations.

This article is devoted to looking for a set of relation types to be leveraged in semantic network created for text classification.

II. Relation Types in Semantic Network

Information systems are often used as a tool to find an answer in response to a query. Let's consider the task of answering to a question in natural language. There are five primary kinds of question in Russian language. Open question is a kind of question which requires clarification [7], for example "who?", "where?", "when?", "how much?". A system able to answer different kinds of

question allows user to make less effort for formulating queries to the system.

Let's imagine that there is a text "Bag stays near the bed", and user asks question "where does the bag stay?". Semantic network used to answer this question [8] may contain the following kinds of relation: "subject-predicate", "place" (Fig. 2).

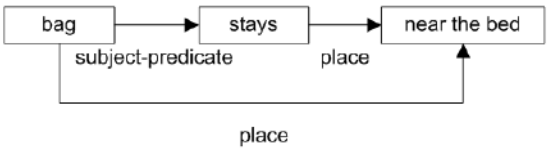

Figure 2. Semantic network used to answer question.

In case of another question other relation types such as "attribute" may be needed. Also, storing synonyms and word forms in network will improve ability to answer questions using the network [8]. Relation types being considered are the following:

1) Word form.
2) Synonym.
3) Subject-predicate.
4) Place.
5) Attribute.

Using these relation types, for the "Black bag stays near the bed" sentence we will get the following more complex semantic network (Fig. 3).

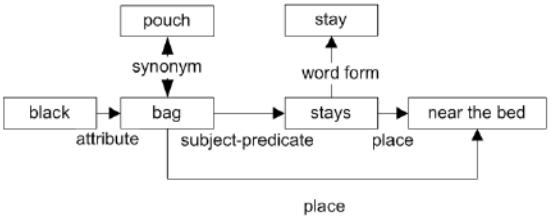

Figure 3. Complex semantic network.

Let's consider using this structure (based on the listed relation types) for text classification. Semantic network is a model which can be used for text classification [9]. Need to note that one of advantages of using this model is that semantic networks correspond to contemporary vision of long-term human memory organization [10].

III. Semantic network generation

Semantic network generation algorithm for a text can be represented as specified below. If a word was already added to the network it is not added the second time.

1) For each meaningful word (e.g. noun, verb, etc.) in each sentence add them to the network.
2) For each meaningful word add stemmed version of the word to the network with "word form" relation type. Porter stemming algorithm can be used to get stemmed version.
3) Load dictionary of synonyms, and for each word in the network find corresponding synonyms. Add found words to the network, connecting them with "synonym" relation type.
4) Find subject(s) and predicate(s) in each sentence. Add them to the network and connect with "subject-predicate" relation.
5) Find words meaning places in the text. Add them to the network and connect with "place" relation.
6) Find words meaning attributes in the text (e.g. adjectives). Add them to the network and connect with "attribute" relation.

This algorithm requires prepared dictionary of synonyms. Subject, predicate, place, attribute words can be found by checking word's part of speech, word form in Russian language (or words order in English language) for simplicity. Also, place can be determined as combination of place pronoun and a noun.

IV. Text Classification

Let's consider a machine learning approach to text classification. To perform text classification, we need to determine a similarity measure for two semantic networks. Similarity S of "network1" to "network2" can be calculated as the following:

$$S = \frac{\sum_{i=1}^{N} Ri}{N} \tag{1}$$

where N is the number of all edges in network1, Ri - similarity of edge number i to network2 (let's call this edge Ei), S - resulting similarity of networks.

The edge Ei has source and target words in network1. If the same words exist in network2 then they may have a $path$ between them in network2. Let's define length of the shortest $path$ as $L2i$.

Ri is calculated as following:

$$Ri = \begin{cases} \frac{1}{L2i}, & \text{if } path \text{ exists,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The idea behind these formulas is that once networks contain the same concepts but connected a bit differently then there is similarity but not 100 percent.

The value of S belongs to range [0;1]. In case of comparing a network with itself S is equal to 1.

Text classification algorithm using semantic network can be represented as following:

1) Determine list of rubrics (i.e. text classes) and texts which will be used for learning. These are texts for which rubric is already known.
2) Define a set of relation types used and a threshold value of similarity required to categorize a text as belonging to a rubric (decision threshold).

3) Concatenate all learning texts for each rubric into single text per rubric. Generate semantic network based on each of texts created. As a result, we have a separate network for each rubric.
4) Create semantic network for classified text and calculate its similarity with networks generated for each rubric.
5) If calculated similarity for one of rubrics is larger then threshold value then the text belongs to that rubric. If threshold is not met for any rubric, or if it is met for several rubrics, then the class is not determined.
6) If there are other texts to classify, go to step 4.

Similarity calculation includes searching for the shortest path between rubric network nodes. This is performed a lot of times for each relation of network created for classified text. So all paths between all nodes of rubric network should be calculated by Floyd–Warshall algorithm. We can make an assumption that words connected with path longer than 2 relations are really not very related and ignore such long chains of relations. Then the algorithm can be limited to look for path with length not longer than 2 which drastically increases performance.

V. Relation Type Sets Comparison

Classification result depends very much on parameters of classification algorithm, in particular on decision threshold amount and relation types used. Contrariwise, algorithms on graphs work relatively slow and count of nodes and edges in the network matters too. So these are primary characteristics being investigated.

Classification algorithm was researched on a set of Russian-language educational texts related to the following 5 disciplines (rubrics): Geometry, Physics, Informatics, Probability Theory, Philosophy, with total size around 2 million characters. The algorithm for calculating similarity includes finding path between all nodes in graph (to improve performance, only path not longer than two edges is considered). The path finding algorithm contains nested loops on nodes and edges so time complexity is $O(n \cdot k^2)$, where n is the number of unique words in text, k - maximal number of relations per word. Average speed of text processing is around 1 Megabyte per 5 minutes.

Based on preliminary check it was decided that it makes sense to use only threshold amounts less than 0.4 because it's very rare that even two texts related to the same rubric would be so much similar. As for relation types, each combination of the 5 relation types was checked:

1) Word form.
2) Synonym.
3) Subject-predicate.
4) Place.
5) Attribute.

The correctness of classification was measured as ratio of correctly classified texts to all texts. In total, 160 experiments were performed. The best results and worst results configurations are displayed in the Table 1. The best result is achieved in case when 3 types of relations are used: Word Form, Subject-Predicate, Attribute. Synonym and Place are not used. It may be related to specific of texts classified: they are mostly technical and synonyms are rarely used in such kind of literature. Also, looks like places written in technical texts are not so different between rubrics as attributes.

On another hand, algorithm based on semantic network containing only Attribute relation shown relatively good results while requesting much less resources. Threshold value in this case is small which means that texts from the same rubric contained small amount of the same attributes, but texts from other rubrics had even less in common.

Semantic network with all five relation types has shown average results, while taking most resources.

Table I
Relation Type Sets Comparison results

Form	Syn.	Subj. -Pr.	Pla ce	At tr.	Correct %	Total Edges	Thres hold
+	-	+	-	+	65	24737	0,14
+	-	+	-	+	57	24737	0,21
-	-	+	+	-	52	17905	0,07
-	-	+	-	+	51	15558	0,07
-	-	-	-	+	46	5366	0,07
-	-	+	-	-	46	10192	0,07
+	-	+	-	+	46	24737	0,28
-	-	-	+	-	42	7745	0,07
+	+	-	+	+	38	30364	0,35
+	-	+	+	+	38	32450	0,14
-	-	-	+	+	37	13111	0,07
-	-	+	-	+	35	15558	0,14
+	-	+	-	-	35	19371	0,21
-	-	+	+	+	34	23271	0,07
-	-	+	+	-	33	17905	0,14
+	+	-	-	+	33	22619	0,35
+	-	+	-	-	33	19371	0,28
...	...	...	...	...	...	...	...
+	+	+	-	+	22	32811	0,21
+	+	+	+	+	22	40524	0,35
-	-	-	-	+	20	5366	0,28
...	...	...	...	...	...	...	...
-	+	+	-	-	1	18266	0,07
-	+	+	-	+	1	23632	0,07
-	+	+	+	-	1	25979	0,07
-	+	+	+	+	1	31345	0,07
+	+	-	-	-	1	17253	0,07
+	+	-	-	-	1	17253	0,14
-	+	-	-	+	0	13440	0,07
-	+	-	+	+	0	21185	0,07
+	+	-	+	-	0	24998	0,07
+	+	-	+	+	0	30364	0,07
+	+	+	-	+	0	32811	0,07
+	+	+	+	-	0	35158	0,07

Need to note that based on all results, this algorithm selects wrong rubric on very rare occasions. The algorithm more likely will not make a decision than make

wrong decision. Based on the formula 1, it makes sense because it's unlikely that common relations would exist in texts related to different topics. Only in such case this algorithm based on semantic network could make a wrong decision.

VI. Conclusion

Semantic networks can be used for different purposes, and relation types used depend on the way in which that network will be used. If the purpose of the network is to answer a question then it is better to have more relations in it. But once another problem is being resolved then more relations doesn't mean better. Semantic network used for text classification have been investigated. The correctness of classification was measured as ratio of correctly classified texts to all texts. Text classification involves several (or more) semantic networks and it may become harmful to have too much relations in each network: they require more processing power, and after including "every possible" relation they become too similar to each other.

Comparison of different sets of relation types has been performed and it shows that combination of 3 relation types (Word Form, Subject-Predicate, Attribute) has maximal percentage of correct results.

Also the investigation has shown that semantic network makes wrong decisions in rare cases so if we use another relation types set and it gives some result, that result is most likely correct.

Using semantic networks for text classification has a drawback: it takes time to analyze each relation type and to find ways in graph. Performance of algorithm analysing each relation in a big semantic network is not very fast. It is possible to combine two or more kinds of semantic networks to improve classification speed. Other possible approaches for improving performance may include caching, hash-tables, selecting most valuable relations based on frequency, simplifying algorithm of finding path on graph. During work on this article, Floyd–Warshall algorithm was modified to find ways not longer than 2 to improve performance.

In general, selecting set of relation types for specific semantic network depends on specific task. There are more and less useful relation sets. It makes sense to perform some preliminary learning and testing of network before using it on real data. Proposed algorithms can be used to perform text classification when performing such tasks as dividing electronic messages on categories, spam filtering, text topic recognition and other.

References

[1] M. Maron. Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, 1961, No. 3, pp. 404 – 417.

[2] Model' strukturnogo predstavleniya tekstovoi informatsii i metod ee tematicheskogo analiza na osnove chastotno-kontekstnoi klassifikatsii. Available at: http://www.dissercat.com/content/model-strukturnogo-predstavleniya-tekstovoi-informatsii-i-metod-ee-tematicheskogo-analiza-na (accessed 2020, Jan).

[3] T.A. Gavrilova and V.F. Khoroshevskii. *Bazy znanii intellektual'nykh sistem* [*Knowledge bases of intelligent systems*], Saint-Petersburg, Piter, 2000. 384 p.

[4] Bazy znanii ekspertnykh sistem. Metody predstavleniya znanii: logicheskie modeli, produktsionnye pravila, freimy, semanticheskie seti. Available at: http://daxnow.narod.ru/index/0-18 (accessed 2020, Jan).

[5] D. R. Rakhimova. Postroenie semanticheskikh otnoshenii v mashinnom perevode [Building semantic relations in machine translation]. *Vestnik KazNU im. al'-Farabi. Seriya «Matematika, mekhanika i informatika»*. [*KazNU messenger named after al'-Farabi. "Mathematics, mechanics and informatics" series.*], 2014, no. 1, pp. 90–101.

[6] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, vol.34, No. 1, 2002, pp. 1–47.

[7] N. Surin. *Ritorika: Uchebnoe posobie* [*Rhetoric: tutorial*], Moskow, MGIU, 2007. 246 p.

[8] V.V. Potaraev. Algoritm primeneniya semanticheskoi seti dlya poiska otveta na vopros [Algorithm of leveraging semantic network to search for answer to a question]. *Komp'yuternye sistemy i seti: Materialy 54-oi nauchnoi konferentsii aspirantov, magistrantov i studentov BGUIR.* [*Computer systems and networks: Materials of 54th scientific conference of postgraduates, undergraduates and students of BSUIR*], 2018, pp. 103–105.

[9] L.V. Serebryanaya, V.V. Potaraev. Metody klassifikatsii tekstovoi informatsii na osnove iskusstvennoi neironnoi i semanticheskoi setei [Methods of textual information classification based on artificial neural and semantic networks]. *Informatika* [*Informatics*], 2016, no. 4, pp. 95–103.

[10] Osnovy iskusstvennogo intellekta. Available at: http://search.rsl.ru/ru/record/01007574162 (accessed 2020, Jan).

Анализ типов отношений в семантической сети, используемой для классификации текста

Потараев В.В.

Современные информационные системы содержат большое количество текстовой информации. Одним из важных способов обработки текстовой информации является классификация текстов. Семантическая сеть является моделью, которая может быть использована для решения различных задач, в том числе для классификации текстовой информации. Существуют другие модели, способные решать эту задачу, и некоторые из них показывают довольно хорошие результаты, но использование семантических сетей имеет такие преимущества, как читабельность и анализ явных семантических отношений между словами. Семантические сети могут иметь различные наборы связей в зависимости от целей, для которых они создаются. В данной работе производится поиск набора типов связей, который можно использовать в семантической сети, создаваемой для классификакации текстов. Осуществляется анализ структуры сетей путём генерации семантических сетей для русскоязычных текстов с использованием различных типов связей. Сгенерированные сети используются для классификации текстов. Классифицируемые тексты были взяты из книг по нескольким техническим дисциплинам. Предложенные алгоритмы могут быть использованы для классификации текста при решении таких задач как разделение электронных сообщений по категориям, фильтрация спама, определение темы текста и при решении других задач.

Received 15.01.2020