The role of phrases in Information Retrieval and related domains

Nikitas N. Karanikolas[1]

[1] Department of Informatics, Technological Educational Institute of Athens,
Ag. Spyridonos street, Aigaleo 12210, Greece
Tel: +302105385882, Fax: +302105910975, E-mail: nnk@teiath.gr

In conventional Information Retrieval (IR), the systems are based on words (actually stems of words) to measure the similarity of documents against the user's information needs. There are also some approaches that extend the model to permit also the use of phrases for expressing the user's information needs. The primary question arising of the introduction of phrases is "what is a phrase?" (what we define as a phrase). A second question is how the phrases can be evaluated together with other (simple) words that compose the user's questions (information needs). More considerable, in our opinion, is the investigation of other usages that phrases can have in other research domains, related with the IR. It is also interesting to investigate the usage of phrases in IR related research domains for applications in Science, Economy, Society and Education. We have elaborated the usage of phrases in document classification and we have got notably promising results. In this paper, we introduce an extension of the mathematical model of similarity, in order to measure the document versus question's (information need's) relevance when questions can be composed by simple words together with phrases. Following, we present the results of our research in the automatic document classification, based exclusively on (automatically selected) phrases. Our, last, contribution in this paper is a conjecture about the way that phrases can be exploited for computer assisted free-text-answers assessment.

Keywords: information retrieval, similarity measure, text classification, computer assisted free-text-answers assessment

1. Introduction

Information Retrieval Systems (Kowalski G., 1997) have been used for the retrieval of documents / information relevant to a submitted query. The queries are submitted either in Natural Language or in a Command Language. Modern Information Retrieval systems (Karanikolas N.N., 2007b) usually present the retrieved documents in decreasing order according to their similarity to the submitted query.

The text classification/categorization is the task of assigning an electronic document to one or more categories (classes), based on its contents. In most cases, text classification is based in some classification function that maps documents to classes. In order to generate the classification function (classifier), we usually apply some supervised learning methodology. The input to the supervised learning method is a training set (a collection of pre-classified, by some external mechanism, documents) and a fixed set of classes (categories) and the output is the classification function. The external mechanism that assigns classes (labels) to the training set documents is named "labeling".

Computer-Assisted Assessment (CAA) of the student essays or the assessment of free text answers to questions is an interesting and recently developing domain. The Natural Language Processing research community has given attention to this long-standing problem. There are some intelligent approaches that include some kind of understanding (of the organization, sentence structure and content) of free text documents (Burstein J. et al, 2001), some pattern-matching techniques (Ming P.Y. et al, 2000) and some Latent Semantic Analysis based approaches (Laham D., 2000; Landauer T. K. et al, 2000).

So far in the Information Retrieval field, the similarity of documents against a user submitted query (that expresses the information needs) is based on tf – idf (term frequency – inverse document frequency) calculations. Terms are (usually) stems of words existing in the document collections. However, the replacement of terms by phrases (n-word adjacent sequences) can improve the system's efficiency (increased precision for the same recall). Next (second) section presents some similarity measure that combines single word terms and phrases.

The learning methodology of a classification system involves a stage that identifies relevant textual characteristics in the documents of the training set and a learning stage in which these characteristics are associated with class labels. The relevant textual characteristics can be structural items, words and phrases. Thus, in case that the classification system is based on phrases, a question arises: which are the relevant phrases to discriminate between classes? This and other similar questions are answered in the third section of this paper.

One of the proposed classification solutions mines (learns) the average class documents. Based on this feature, computer assisted assessment of a free text answer can be decomposed to a number of similarity measurements. For example, if we have a number of questions (classes), a number of positive (correct) answers for each class (question) and a number of negative (incorrect) answers for each class then we can mine two average class documents ("positive average class document" and "negative average class document") for each class. Consequently, the assessment of a given answer for a given question (class) can be decomposed to a) the similarity of the assessed answer with the positive average class document of the given question, b) the average similarity of the assessed answer with all positive average class documents of the rest questions (classes) and c) the similarity of the assessed answer with the negative average class document of the given question.

2. Relevance ranking based on phrases

The problem of similarity for a document against a submitted query, also known as Document versus Query similarity or nearest neighbors has been the field of continuing research for more than 20 years (Smeaton A.F. & van Reijsbergen C.J., 1981). Many similarity functions / measures and algorithms eliminating the query-document comparisons have been proposed (Bentley J.L. et al, 1980; Lucarella D., 1988; Borlund & Ingweren, 1998).

In the popular Vector Space Model (VSM) a data set of n unique terms is specified, called the index terms (or keywords) of the document collection, and every document can be represented by a vector:

$(T_1, T_2, \ldots, T_n)$

where T_i=1, if the index term i is present in the document, and 0 otherwise.

A query can be represented in the same manner. The document and query vectors can be envisioned as an n-dimensional vector space. A vector matching operation, based on the cosine correlation used to measure the cosine of the angle between vectors, can be used to compute the similarity. Hence, the following equation (Lucarella D., 1988) gives us a well-known method to measure the similarity of document D_i against query Q:

$$S(D_i,Q)=\frac{\sum_{j=1}^{n} q_j t_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \cdot \sum_{j-1}^{n} t_{ij}^2}}=\frac{\sum_{j=1}^{n} q_j t_{ij}}{L_Q \cdot L_{D_i}} \qquad (1)$$

where n is the number of index terms used in the collection, t_{ij} is the weight of term j in document D_i and q_j is the weight of term j in the query.

The following two equations can be used to measure the terms t_{ij} and q_j:

$$t_{ij} = 0.5 + 0.5 \cdot \frac{F_{ij}}{\max F_i} \qquad (2)$$

$$q_j = \log_2\left(\frac{N}{DOCFREQ_j}\right) \qquad (3)$$

where F_{ij} is the frequency of term j in document D_i, $\max F_i$ is the maximum frequency of the terms in document D_i, N is the number of documents in the collection and $DOCFREQ_j$ is the number of documents that include the index term j.

According to (1) some kind of the “document length” is given by:

$$L_{D_i} = \sqrt{\sum_{j-1}^{n} t_{ij}^2} \qquad (4)$$

We have introduced (Karanikolas N.N., 2007a) an alternative calculation of the “document length”:

$$L_{D_i} = \ln\left(\sum_{j=1}^{n} t_{ij}^2 + e - 1\right) \qquad (5)$$

This, later calculation of the “document length”, resolves the problem of “preference” to short documents against longer ones that otherwise is present.

So far we have not considered the use of phrases for the calculation of Document versus Query similarity. For the exploitation of phrases we suggest the following contribution of each phrase in the nominator of equation (1):

$$c \cdot q_{\{\mathrm{A..\Delta}\}} \cdot \left(0.5 + 0.5 \cdot \frac{F_{i\{\mathrm{A..\Delta}\}}}{\max F_i}\right) + B \cdot \sum_{x \in \{\mathrm{A..\Delta}\}} q_x \cdot t_{ix} \qquad (6)$$

This is the calculation of a phrase’s contribution and replaces $q_j t_{ij}$ (the contribution used otherwise – for simple query words). The following are the new terms introduced by (6):

- $t_{ix} = 0.5 + 0.5 \cdot \frac{F_{ix}}{\max F_i}$
- $q_x = \log_2 \left(\frac{N}{DOCFREQ_x} \right)$ where $x \in \{\mathrm{A}..\Delta\}$
- $q_{\{\mathrm{A}..\Delta\}} = \max\limits_{x \in \{\mathrm{A}..\Delta\}} (q_x)$
- $B = b / |\{\mathrm{A}..\Delta\}|$
- $\max F_i$ is the maximum frequency of simple term (not phrase) appearance in document i
- $F_{i\{\mathrm{A}..\Delta\}}$ is the frequency of the phrase {A..Δ} in document i
- c and b are constants determining what is the contribution of the phrase and what is the contribution of the phrase constituents (words).

For the denominator of equation (1), the calculation of LD$_i$ remains unchanged. However, for the calculation of LQ, its calculation considers simple terms (q_j) and phrases ($q_{\{\mathrm{A}..\Delta\}}$) and none of the phrase constituents (q_x where $x \in \{\mathrm{A}..\Delta\}$).

In other words, every phrase (compound term) of the query is taken as a simple term but with increased weight. The standard weight is given by:

$$c \cdot q_{\{\mathrm{A}..\Delta\}} \cdot \left(0.5 + 0.5 \cdot \frac{F_{i\{\mathrm{A}..\Delta\}}}{\max F_i} \right)$$

and the increased weight result by the add up of:

$$B \cdot \sum_{x \in \{\mathrm{A}..\Delta\}} q_x \cdot t_{ix}$$
.

Our motivation for using increased weightiness is double:

- The documents that contain the query phrase (compound term) gain a heavy bounty.
- The documents that do not contain the query phrase but contain some of the phrase constituents (words) gain some (less heavy) bounty. (In this way the documents that include only phrase constituents appear at the end of the results list.)

It remains to explain which is the role of the parameters c and B and how this parameters can be configured. The parameter c determines the contribution of the phrase (as an unbreakable whole). The parameter B determines which is the contribution of each of the phrase constituents (words). Although parameter B is a necessary part of (6), it is not of interest to the user. The user should be able to determine what is the contribution of all phrase constituents (words). This contribution is determined by the parameter b. The parameter B results as a division with nominator the parameter b and denominator the number of words that constitute the phrase.

The user's intervention can be based on two handlers. First handler should define what is "the weight of a phrase against the weight of a simple term". This handler determines the sum of parameters c and b. The second handler should define what is "the constituents' contribution in the weight of phrase". This second handler determines the value of $b/(c+b)$. For example if the user defines that "the weight of a phrase against the weight of a simple term" is 1.32 and "the constituents' contribution in the weight of phrase" is 0.25, then (c+b)=1.32 and (b/(c+b))=0.25. Consequently, b=0.33 and c=0.99.

The range of values of the first handler ("the weight of a phrase against the weight of a simple term"), in our opinion, should be from 1.0 to 3.0. The range of values of the second handler ("the constituents' contribution in the weight of phrase"), in our opinion, should be from 0.0 to 0.5. These ranges are not definite but they can be used as a first approach.

3. Classification based on phrases

Phrases can be used as features of texts and based on their existence (or not) to mine text classification rules (Karanikolas N.N. and Skourlas C., 2002; Karanikolas N.N. et al, 2003). A significant factor for the success of the mined classification rules is the appropriate selection of phrases. In other words, we should be able to select only the phrases that are able to discriminate between classes. We call this set of phrases (that we use to discriminate between classes) "Authority List". In our research, we have concluded to the following rule for extracting the Authority List (Karanikolas N. N. and Skourlas C., 2004):

> Given a collection of documents subdivided into classes and a window width, find all key-phrases that occur frequently enough in one or few classes but do not occur frequently enough in other classes. A key-phrase is considered as existent in a document if all of its constituents (words) occur in the document, in the same order as in the key-phrase, and the distance of the first to the last constituent (word) in the document is not greater than the window width.

So far, we have introduced two methods for classifying new unclassified documents (Karanikolas N. N. and Skourlas C., 2005; Karanikolas N. N. and Skourlas C., 2006). The first one is based on the similarity of the new document with the "Average class document" (mined through the training phase) of each class. The second method ("All documents") uses the similarity of the new document with every document in the training set in order to calculate the average similarity of the new document with each class. The second method has slightly better results than the first one but it is more expensive in computer power.

4. Text assessment based on phrases

The similarity of a new document (a provided answer by some student in an examination) with a given class (the question for which the user provides the answer) and the similarities of the new document (answer) with all other classes (the rest questions in the pool of questions) can constitute two (2) features for the evaluation (estimation) if the given answer (the new document) is a correct answer for the given question (class). These two (2) features are calculated on the strength of the positive instances (correct answers) of the training set. The similarity of a new document (answer) with any class (any question) can be based either in the "Average class document" method or in the "All documents" method (see previous section for these methods).

In order to exploit (turn to advantage) the negative instances of classes (wrong answers given for the questions) we can do the following:

1. create the global Authority List by using only the positive instances (correct answers) of the training set. As a side effect, for each class, it is created the class specific Authority sub-List (the subset of phrases of the global Authority List that occur frequently enough in the class documents – from now on Class_AL_Valid)
2. create the global Authority List by using both positive and negative instances (correct and incorrect answers) of the training set. The created, as a side effect, class specific Authority sub-Lists contain phrases from both positive and negative (All) instances of the corresponding class (from now on Class_AL_All)
3. Make a loop over the classes and for the examined each time class, make a "subtraction" (between Class_AL_All and Class_AL_Valid) in order to find the key-phrases that characterize only the negative instances of the class (from now on Class_AL_Invalid)

In this way, the features that are used for the calculation of the overall similarity of a new document (provided answer) against a class (question for which the answer is provided) are increased to the following three (the last feature is the newer one):

- simple similarity of the new document with the class (this feature works positively for the overall similarity),
- average similarity of the new document with the rest classes (this feature works negatively for the overall similarity),
- calculation (with the "Average class document" method) of the simple similarity of the new document with the Class_AL_Invalid of the question (this feature works negatively for the overall similarity).

5. Conclusions

We have considered ways to incorporate phrases (sequences of words) to the methods used in Natural Language Processing (NLP) systems. We have considered the use of phrases in conventional Information Retrieval (IR), Text Classification and Free-text answer assessment. We have given emphasis to the use of phrases in IR systems, because this is our main contribution in this paper. The main aspect we have considered is how phrases can be evaluated together with other (simple) words that compose the user's questions (information needs). We have also briefly discussed the usage of phrases for text classification. For more details about phrases and text classification, the interested reader can consult our previous work (see references provided in the relevant section of paper). For the third sub-domain of NLP that we have considered (Free-text answer assessment) we have presented our suggestions for promising features that can be used for judging if an answer is a correct answer for a given question. These features constitute our conjecture and have not evaluated yet.

References

[1] Bentley, J. L., Weide, B.W. and Yao, A. C., “Optimal expected time algorithms for closest point problems”, *ACM Trans. Math. Software*, vol. 6, pp. 563-580, 1980.

[2] Borlund and Ingweren, “Measure of relative relevance and ranked half life: performance indicators for interactive IR”, *Proc. SIGIR’98*, pp. 324-331, 1998.

[3] Burstein, J., Leacock, C., & Swartz, R., “Automated evaluation of essay and short answers”, In M. Danson (Ed.), *Proc. Fifth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK, 2001.

[4] Karanikolas N. N. and Skourlas C., „Automatic Diagnosis Classification of patient discharge letters”, *MIE’2002: XVIIth International Congress of the European Federation for Medical Informatics*, Budapest, Hungary, August, 2002. IOS Press, ISBN: 1-58603-279-8, ISSN: 0926-9630.

[5] Karanikolas N. N., Skourlas C., Christopoulou A. and Alevizos T., “Medical Text Classification based on Text Retrieval techniques”, *Proc. MEDINF 2003: 1st International Conference on Medical Informatics & Engineering*, Craiova, Romania, October 2003.

[6] Karanikolas N. N. and Skourlas C., „Key-Phrase Extraction for Classification”, *Proc. MEDICON and HEALTH TELEMATICS 2004: X Mediterranean Conference on Medical and Biological Engineering*, Ischia, Italy, July 31 - August 5, 2004.

[7] Karanikolas N. N. and Skourlas C., “Naive Rule Induction for Text Classification based on Key-Phrases”, *Proc. of the 6TH International Conference on Data Mining, Text Mining and their Business Applications*, pp. 175-182, Skiathos, Greece, May 2005. WIT Transactions on Information and Communication Technologies, vol. 35, Data Mining VI: Data Mining, Text Mining and their Business Applications, WIT Press, ISBN 1-84564-017-9, ISSN 1764-4463.

[8] Karanikolas N. N. and Skourlas C., “Text Classification: Forming Candidate Key-Phrases from Existing Shorter Ones”, *FACTA UNIVERSITATIS Series: Electronics and Energetics*, ISSN 0353-3670, vol. 19, no. 3, 2006.

[9] Karanikolas N. N., “The measurement of similarity in stock data documents collections”, *eRA-2. 2nd Conference for the contribution of Information Technology to Science, Economy, Society and Education*, Athens, Greece, September 22-23, 2007.

[10] Karanikolas N. N., “Low cost, cross-language and cross-platform Information Retrieval and Documentation tools”, *Journal of Computing and Information Technology (CIT)*, ISSN: 1330-1136, vol. 15, no. 1, 2007.

[11] Kowalski G: “Information Retrieval Systems. Theory and Implementation”, 1st edn (Printed in the USA; Kluwer Academic Publishers), ISBN 0-7923-9899-8, 1997.

[12] Landauer, T. K., Laham, D., & Foltz, P. W., “The Intelligent Essay Assessor”. In M. A. Hearst (Ed.), The debate on automated essay grading, IEEE Intelligent systems, 27–31, September/October, 2000.

[13] Laham, D., “Automated content assessment of text using Latent Semantic Analysis to simulate human cognition”, PhD Dissertation, University of Colorado, Boulder, 2000.

[14] Lucarella, D., “A document retrieval system based on nearest neighbor searching”, *Journal of Information Science*, vol.14, pp. 25-33, 1988.

[15] Ming, P.Y., Mikhailov, A.A., & Kuan, T.L., “Intelligent essay marking system”, In C. Cheers (Ed.), Learners Together, NgeeANN Polytechnic, Singapore, Feb. 2000.

[16] Smeaton A.F. and van Reijsbergen C.J., “The Nearest Neighbour problem in Information Retrieval. An algorithm using upperbounds”, *ACM SIGIR Forum 16*, pp.83-87, 1981.