Geometrical approach for modeling semantics in linguistics

© Milan Gudába[1], Stanislav Horal[1], Ladislav Izakovič[1],
Michaela Kalinová[1] and Václav Snášel[2]

[1]Department of informatics, FPV, University of Saint Cyril and Methodius,
Nám. J. Herdu 2, 917 01, Trnava, Slovakia

[2]Department of informatics, FEI, VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33, Ostrava-Poruba, Czech republic

gudaba@gmail.com, stanislav.horal@ucm.sk, izakovil@ucm.sk, michaela.kalinova@ucm.sk,
vaclav.snasel@vsb.cz

Abstract

The information is at the present time often saved and available in electronic form. With still increasing quantity of accessible, most frequently text information, the need of organization of these data is raising. The problem of fast and effective information retrieval occurs very often. In this contribution we describe the method for creating word vector space and using NOT operation for more effective acquirement of relevant documents.

1 Introduction

High volume of text documents and the rate of growth these documents requires finding new approaches in linguistics and in areas related with information retrieval (IR). These new approaches are based on principles which are derived from natural sciences.

Noticeable expansion of geometrics was motivated by Descartes ideas and by establishment of the coordinate system which permit interconnection between geometrics and algebra [12, 13].

In IR was geometrical methods enveloped in the form of the vector model. Another meaningful step in geometrical understanding of the world was interconnection of geometrics and logics. This connection was made on the ground of the quantum physics [9, 12].

A huge amount of multimedia data is at the present time coupled with expansion of information technologies. Among these data we can include especially text, image and acoustic documents. The set of text documents we will be consider as an input area. Almost in all well-known information retrieval systems occurs the morphological part. By the help of this part we can, by using stop-list, remove non-semantic words from documents, and semantic significant words convert to the basic form. In this way, we specify terms which after evaluation make vector in the space of concepts. This vector is then used for documents identification from the point of view his content [16].

The application of the method based on combination of the vector model and Boolean logic appears as a suitable way for creating the model of natural language. In this contribution we present constructions in the vector space based on the standard linear algebra, and also some examples of using the vector negation for separation meanings of ambiguous words. In quantum logic arbitrary sets are substituted by linear subspaces of the vector space and union, intersection and complement are substituted by the vector sum, intersection and orthogonal complements of these subspaces.

A useful tool for information retrieving and processing is latent semantic analysis – LSA. This method, which is based on singular value decomposition (SVD), we can use for improving access to desirable documents. We regard the factor of greatest singular values *k*. LSA has geometrical representation, in which objects (e.g. documents and terms) are distributed in the low-dimensional space. As an example we can use term-document matrix, in which rows represent terms and columns of matrix represent documents. Nonzero values in matrix signify that corresponding documents include required terms. This vector space model was described by Salton [10].

In the first part of our contribution is described representation of word meanings in the vector space. Second part is focused on operations in the word space. Next chapter is devoted to basic knowledge from Boolean logic. In the last part we mentioned theoretical knowledge applied in the process of searching word meanings.

2 Word meaning representation in vector space

Vector space could be understood as a set of points, in which each point of the space is defined by the list of coordinates [5]. Two points can be accountable by adding theirs coordinates and each point can be multiplied by the scalar (in this paper scalars are real numbers, therefore all our vector spaces are "real" vector spaces). The first linguistic examples of vector spaces were developed for information retrieval [10]. By accounting occurrences of each word in the each document we get term-document matrix. Each couple i,j in matrix indicates number how many times was the word w_i occurred in the document D_j. Then rows of the matrix can be understood as word-vectors. Dimensions of this vector space (number of coordinates given to each word) are therefore equal to the number of documents in collection. *Document vectors* are generated by calculating (weighted) sum of word-vectors of words occurred in given document.

Similar techniques are used in information retrieval for determination similarity relation between words and documents. Similarity can be determined by calculating cosine of the angle between two vectors [10], where w_i, d_i are coordinates of the vectors w and d, $w \cdot d$ is inner product w and d. $\|w\|$ is length of the vector w [5].

$$sim(w,d) = \frac{\sum w_i d_i}{\sqrt{\sum w_i^2}\sqrt{\sum d_i^2}} = \frac{w \cdot d}{\|w\|\|d\|}$$

The calculus is further simplified by normalization of all vectors to the same length, consequently then the cosine similarity is equal with euclidean inner product. This is standard method which avoids add great weight of consequence to frequent words or large documents. Normalized vectors was used in all models and experiments described in this contribution.

This structure can by used for determination similarities between pairs of words – two words will have high similarity, if they will be situated in same documents and only seldom is one word occurred without another. Some words are combined into combined query statements by using the commutative vector sum.

The *term-document matrices* are typically very sparse. Information could be concentrated in low number of dimensions when we use singular values from decomposition, transformation each word into *n*-dimensional subspace. This guarantees method of least squares. Each word is then represented by using *n* most significant latent variables. This process is called *latent semantic analysis* – LSA [6]. Especially for these purposes of determining semantic similarity between words was made by Schütze one variant of LSA [11]. Instead of using documents as columns in matrix, there were used *content-bearing* words. Consequently, in our case is vector of the word *koruna* (crown) defined upon that, that it was frequently occurred by the words *cena* (price) and *mena* (currency). This method is convenient for semantic tasks, like is clustering words according to similar meaning and determination disambiguation of words [7, 8].

3 Logical operations in word space

At investigation dependencies of words in the word space we can use logical operations, mainly primarily negation in relations of orthogonality and disjunction in relations of the vector sum of subspaces.

3.1 Vector negation

We would like to model meaning of expression „*koruna* NOT *klenot*" (*koruna* – crown as a coin, *klenot* – crown as a jewel) in a such way, that system will be able to awake, that we are interested in finances, but not about meaning of the word *koruna* in the sense of the jewel. Therefore we need to find aspects of meaning the word *koruna* which are different from the word *koruna* as a jewel, and have no relation to this word. Word meanings have not interrelationship if they have not any common marks. Document is considered as irrelevant for user if the inner product with user query is equal to zero, when query vector and document vector are orthogonal [5].

Definition 1. *Two words a and b are considered as irrelevant to each other, if their vectors are orthogonal, i.e. a a b are irrelevant to each other, if* $a \cdot b = 0$ [15].

Definition 2. *Let V is a vector space with inner product. Then we could define for vector subspace A⊆V orthogonal subspace* $A^{\perp}$ [15]

$$A^{\perp} \equiv \{v \in V: \forall a \in A, a \cdot v = 0\}.$$

If A and B are subspaces of the space V, then NOT B represent $B^{\perp}$ and A NOT B represent projection A into $B^{\perp}$. When *a*, *b* belongs to V, then *a* NOT *b* represent projection *a* into $\langle b\rangle^{\perp}$, where $\langle b\rangle$ is the subspace $\{\lambda b : \lambda \in \mathbb{R}\}$.

These definitions can be used for realization calculations with vectors in vector space. We apply standard method of projection.

Theorem 1. *Let a, b are subsets of V. Then a NOT b is represented by vector*

$$a\ \ NOT\ \ b = a - \frac{a \cdot b}{|b|^2} b\,,$$

where $|b|^2 = b \cdot b$ *is length of the vector* b [15].

Example: When we are doing inner product with b, then we obtain

$$(a\ \ NOT\ \ b) \cdot b = \left(a - \frac{a \cdot b}{|b|^2} \cdot b \right) = a \cdot b - \frac{(a \cdot b)(b \cdot b)}{b \cdot b} = 0$$

This proves that *a* NOT *b* and b are orthogonal, therefore vector *a NOT b* is certainly part of *a*, that is irrelevant to *b* (Definition 1), as we required.

When we have normalized vectors, then Theorem 1 has following form

$$a\ \ NOT\ \ b = a - (a \cdot b) b\,.$$

For the purpose of finding expressions or documents, which correspondent to *a NOT b,* is not important for each candidate from a as well as *b* consequently determine certain differences. Theorem 1 shows that finding similarity between another vector and *a NOT b* is simple calculus of inner product.

4 Quantum logic and vector space

With concept of quantum logic we met for the first time in the theory of the quantum mechanic, which was presented by Birkhoff and von Neumann (1936) [2]. From the set theory is known, that if we have sets A and B and the element $a \in A\ \ or\, a \in B$, then also union of these sets $C = A \cup B$ will contain this element. But the quantum logic does not describe A and B as sets, but as subspaces of the vector space.

The structure of the quantum logic is simple and we can obtain it by substitution of sets and subspaces by vector spaces and subspaces [2]. Points in quantum mechanics are represented as subspaces of the vector space V. In this connection we can consider the collection of subspaces L(V) in the vector space V. The lower bound of $A, B \in L(V)$ is the biggest element $C \in L(V)$, where $C \subseteq A$ and $C \subseteq B$, what is exactly conjunction of $A \cap B$. The upper bound of A and B is the smallest subspace $D \in L(V)$, where $A \subseteq D$ and $B \subseteq D$. These two operations give partially formed set L(V), which is structure of the lattice. If we work in the area of scalar product, we can define for each subspace $A \in L(V)$ its (special) orthogonal complement $A^{\perp}$. So we have three operations which we use in collection of L(V) and are defined as follows [2]:

Conjunction	A AND B = $A \cap B$
Disjunction	A OR B = $A + B$
Negation	NOT A = $A^{\perp}$

It is simple to prove, that these three operations on L(V) are sufficed to realize any necessary relations ($A + A^{\perp} = V, A \cap A^{\perp} = 0$) and to define logic on L(V).

The important piece of knowledge is also that every subspace $A \in L(V)$ can be identified (by using scalar product) with special projective map $P_A : V \to A$ and through this bijection is logic of subspaces L(V) equivalent to logic of projection mapping in the vector space V.

The quantum logic is distinguished from Boolean logic at least in two properties: quantum logic is not distributive as well as commutative.

The disjunction in set theory can be modeled as union of sets, which corresponds in linear algebra to the vector sum of subspaces, where A+B is the smallest subspace of V containing A as well as B.

For determination of similarity between arbitrary objects is necessary to define some function σ: D×D → R. These functions assign a real number to pair of objects o_i, o_j from their domain area D. This formula will be a measure of similarity relation of objects, which must satisfy following requests:

1. $\sigma\ (o_i, o_j) \geq 0$
2. $\sigma\ (o_i, o_j) = \sigma\ (o_j, o_i)$, i.e. remaining of symmetry
3. when $o_i = o_j$, than $\sigma\ (o_j, o_i) = \max \sigma\ (o_k, o_l)$; for $\forall\ o_k, o_l \in D$

Definition 3. *Let terms b_1 ... b_n ∈ V. Term b_1 OR ... OR b_n is represented by thesubspace* [15]

$$B = \{\lambda_1 b_1 + \ldots + \lambda_n b_n : \lambda_i \in R\}.$$

The search of similarity relation between an individual term *a* and a general subspace B, is more complicated as a search of similarity relation between individual terms.

From the look of quantum physic, we can use P_B to measure a probability that any element was found in some state [12]. The value $a \cdot P_B(a)$ is interpreted as a measured probability. For our purposes we define probability with following relation

$$sim(a, B) = a \cdot P_B(a),$$

where probability is given by scalar product a with projection a to the subspace B, from which we calculate value of term a lying in subspace B. Problematic similarity relation was searched from various looks, see [1],[12].

In practice, if the set $\{b_j\}$ is orthonormal then it is not correct only to calculate $sim(a,b_j)$ for every vector b_j in order. For obtaining an orthonormal base $\{\tilde{b}_j\}$ for subspace B is convenient firstly construct orthonormal base for B by in practice used Gram-Schmidt method [5].

$$P_B(a) = \sum_j \left(a \cdot \tilde{b}_j\right)\tilde{b}_j$$

Consequently, we can write

$$sim(a,B) = \sum_j \left(a \cdot \tilde{b}_j\right).$$

For enumeration $sim(a,B)$ we need to calculate a scalar product a with every vector $\tilde{b}_j$. This similarity relation is more difficult to calculate as in Theorem 1. The result which we reached by comparing every document *a NOT b* using only one operation - scalar product, is the loss for disjunction, although how we will show later, but is desirable for negated disjunction.

5 Using the negation for search meanings

In this part is presented an introductory example of vector connections which demonstrate usage of vector negation and vector conjunction, vector disjunction and negation together for finding vectors which represent different meanings of ambiguous words. We describe shortly a document of obtained experiment. It shows that vector negation has smart contribution contrary of classic Boolean method which was described in paper [15].

Our word space was constructed from 28 articles written in year 2006 which was obtained from the Internet. The total number of acquired words was 5260. The collection of articles was focused on economy, culture, sport, health and science and from every sphere was processed at least two articles. Documents which concern meanings of the word *koruna* (crown as coin) are marked as D13, D14, D15. On the other hand, documents related to *koruna* (crown as jewel) are marked D10, D11, D

Over data source was created parser which separated individual words from the text. By using morphological analyzer [4] was consequential constructed a list of terms. We assume that in articles occurred meanings of ambiguous words. For example, word *koruna* (crown as coin) is used more frequently in economic context as in context common with jewels. For testing the effectiveness of our operator of negation we will try to find less common meanings of chosen words which are related with prevailing expression.

Table 1. Example of the influence of the factor k on the relevance of documents.

koruna

k=2		k=8		k=15	
D14	1	D13	0,773	D13	0,657
D15	0,999	D15	0,647	D23	0,563
D10	0,999	D14	0,643	D18	0,478
D13	0,998	D12	0,637	D10	0,432
D24	0,967	D10	0,597	D15	0,413
D28	0,96	D11	0,592	D14	0,384
D25	0,958	D22	0,561	D11	0,349
D27	0,956	D24	0,363	D22	0,327
D11	0,952	D16	0,336		
D26	0,946	D18	0,328		
D22	0,935	D23	0,324		
D07	0,928				
D12	0,912				
D08	0,908				
D04	0,905				
D16	0,902				
D03	0,885				
D21	0,879				
D06	0,878				
D05	0,85				
D20	0,85				
D17	0,796				
D18	0,751				
D23	0,716				
D09	0,603				
D19	0,585				

The present experiments represent a calculation of similarity in relationship term-document for different values of k in area <2, 15>. Here we illustrate results for factor value *k=2, k=8 and k=15.*

The data in Table 1 represent that vector negation is very effective for selection of relevant documents which correspond to the required word *koruna* (crown) and left out the word *klenot* (jewel).

LSI regards k greatest singular values. The choice of k have to be enough small for obtaining faster access to documents, but enough great for adequate interception of structure of the corpus.

Table 2. Example of the influence of the operation NOT and the influence of the factor k on the relevance of documents.

koruna NOT klenot

k=2		k=8		k=15	
D14	1	D13	0,806	D13	0,621
D13	0,999	D15	0,677	D23	0,57
D15	0,999	D14	0,671	D18	0,55
D10	0,998	D22	0,601	D15	0,439
D24	0,96	D10	0,582	D10	0,41
D25	0,951	D18	0,386	D14	0,4
D27	0,95	D24	0,374		
D28	0,95	D23	0,374		
D26	0,94	D16	0,353		
D07	0,927	D20	0,328		
D22	0,924				
D08	0,905				
D04	0,905				
D16	0,883				
D06	0,881				
D03	0,88				
D21	0,867				
D05	0,844				
D20	0,823				
D17	0,768				
D18	0,732				
D23	0,696				
D09	0,605				
D19	0,542				

We realized a decomposition of matrix A for different values of k. The most relevant documents were obtained for value of factor k=15. On the contrary, for the very small k=2 is obtained great volume of documents. It reduces their relevance in regard to required document.

The vector negation and disjunction can be combined with selection of some searched query from areas of documents. We do not negate only one argument, but several. If the user determines that he wants documents related with a but not with $b_1,b_2,...,b_n$, it will be interpreted (without next indication) that he wants only documents witch are not related with unwanted terms b_i. In this way, the next expression

$$a\ AND\ (NOT\ b_1)\ AND\ (NOT\ b_2)\ ...\ AND\ (NOT\ b_n)$$

will pass into form

$$a\ NOT\ (b_1\ OR\ b_2\ ...\ OR\ b_n).$$

By using Definition 3 we will form a disjunction b_1 *OR* b_2 ... *OR* b_n as vector subspace $B = \{\lambda_1 b_1 + ... + \lambda_n b_n,\ \lambda_i \in R\}$. This term can be transformed on definite vector which is orthogonal to all irrelevant arguments $\{b_j\}$. This vector will be $a - P_B(a)$, where P_B is projection on the subspace B as in Theorem 1. This implies that calculus of the similarity between all vectors with term a *NOT* $(b_1$ *OR* b_2 ... *OR* $b_n)$ is the same as simple scalar product which has the same computing effectiveness as the Theorem 1. This technique is assigned to systematic reduction of irrelevant terms.

6 Conclusion and future work

The negation in the vector space is suitable tool for dimension reduction of required documents. The specification of searched documents can be realized by using several disjunction operations in query. Our word space consisted of 5260 words. More relevant results can be obtained by application this method for largest document database.

Actual experiments represent calculus of similarity in the relation term-document. By construction of the vector space model we have represented individual occurrences in documents through Boolean function, i.e. we have expressed attendance if you like absence of the given term in the document.

Contemporary experiments will be in future carried out over lexical database of words and the word connections in the WordNet, where are recorded relative lexical and semantic relations between individual contained words or concepts.

In the next experiments we target our effort to calculation of similarity in relation term-term by various forms of representation the weight of individual term in the document.

References

[1] Bělohlávek, R., Snášel, V.: Podobnost a její modelování, *Znalosti 2004*, p 309-316

[2] Birkhoff G., von Neumann J. The logic of quantum mechanics. *Annals of Mathematics, 37, 823–843.* 1936.

[3] Gärdenfors P. Conceptual Spaces: The Geometry of Thougt. *MIT Press 2004.* pp 307.

[4] Horal S., Kalinová M., Kostolanský E.: Morfologická analýza slovenčiny - Analýza flexných tvarov. Conference Proceedings *Informatika 2005*. Bratislava 2005.

[5] Jänich K. *Linear algebra*. Undergraduate Texts in Mathematics. 1994, Springer-Verlag.

[6] Landauer T., Dumais S. A solution to plato's problem: *The latent semantic analysis theory of acquisition. Psychological Review, 104(2), 211–240.* 1997.

[7] Moravec P., Pokorný J., Snášel V. *Using BFA with WordNet Ontology Based Model for Web Retrieval.* Yaoundé 27.11.2005-1.12.2005. In: CHBEIR, Richard, DIPANDA, Albert, YÉTONGNON, Kokou (ed.) SITIS 2005. Dijon : University of Bourgogne, 2005, p. 254-259.

[8] Moravec P., Pokorný J., Snášel V. *WordNet Ontology Based Model for Web Retrieval.* IEEE WIRI 2005. Japan Tokyo, p. 220-225.

[9] Pokorný J., Snášel V., Kopecký M. *Dokumentografické informační systémy.* Karolinum, Skriptum MFF UK Praha, 2005, pp 184.

[10] Salton G., McGill M. *Introduction to modern information retrieval.* McGraw-Hill, 1983, New York, NY.

[11] Schütze H. Automatic word sense discrimination. *Computational Linguistics, 24(1), 97–124.* 1998.

[12] Tversky A.: Features of similarity. *Psych. Rev. 84 (1977)*, 327–352.

[13] Van Rijsbergen K. *The Geometry of Information retrieval.* Cambridge University Press. 2004, pp 150.

[14] Widdows D. Geometry and Meaning. *CLSI Lecture Notes, No. 172*. Stanford, California, 2004, pp 320.

[15] Widdows D., Peters S. Word Vectors and Quantum Logic Experiments with negation and disjunction, *Appeared in Mathematics of Language 8*, Indiana, June 2003, p. 141-154. Proceedings of Mathematics of Language 8, 2003

[16] Zee A.: Quantum Field Theory. Princeton University Press. 2003.