Uselessness of simple co-occurrence measures for IF&IR — a linguistic point of view.

Dominik Kuropka
Hasso Plattner Institute for IT-Systems Engineering
at the University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
dominik.kuropka@hpi.uni-potsdam.de

Abstract

This paper gives a short motivation for the use of similarities of term pairs to improve information filtering and retrieval methods. Further some simple, but often used co-occurrence based measures for the estimation of similarity values of term pairs are presented. The results of these measures are evaluated against the claims on term similarities, which arise from linguistic aspects (inflection, hyponymy, meronymy, composition and word group) between the involved terms. The finding of this paper is the uselessness of simple co-occurrence based measures for the improvement of information filtering und retrieval approaches.

1. Introduction

The objective of Information Filtering (IF) and Information Retrieval (IR) is the selection of relevant documents from a huge pool of electronic documents. While IF selects documents from a dynamic stream of documents using some kind of (static) profiles, IR selects relevant documents from a static set of documents basing on a specified (ad hoc) query [BeCr1992]. Because of the high similarity of these tasks, most concepts and models for IF or IR can be adapted for both tasks. [Kuro2004]

Before processing documents by IF or IR, documents are usually split up into terms (in most cases single words). Classic IF and IR approaches like the Vector Space Model [Salt1968] or the Binary Independence Retrieval [RoJo1976] are characterized by the assumption of *term independency* [BaRi1999]. Term independency stands for the fact that two different terms (e.g. ‘car’ and ‘wheel’) are assumed to have no relationship to each other. Because of this escapist assumption of the classical approaches, a several new approaches has been published in the last three decades. These approaches try to overcome the assumption of term independency. Most of these approaches model the dependency between terms by the use simple measures for *co-occurrence of terms* within a given document base to specify the degree of dependency without need for human intervention. Popular models for example are: the Generalized Vector

Witold Abramowicz (ed.), Business Information Systems, Proceedings of BIS 2005, Poznań, Poland

Space Model [WZRW1987], and the family of Spreading Activation Neuronal Network models [Belew1989] [Kwok1989] [WiHi1991].

In the beginning the new models looked very encouraging. Thus, it was a big disappointment that the new models using simple term co-occurrence measures to determine term dependencies are not significantly better than the classic models [BaRi1999]. This is the reason why newer approaches (e.g. the Latent Semantic Index model [FDDL1988], the Language Model [SoCr1999] and most of its derivates) use more sophisticated statistical methods or adaptive approaches (e.g. Fuzzy Set Model [OgMK1991], COSIMIR [Mand2001]) or ontology-based approaches (e.g. (enhanced) Topic-based Vector Space Model [BeKu2003] [Kuro2004]) for term dependency determination.

While the topic of interest in the IF and IR community has moved away from simple co-occurrence based approaches, the reason for the failure of approaches basing on simple co-occurrence measures has not been worked out so far. However, it is beneficial to know, why things went wrong to avoid similar mistakes in the future. For this reason in this paper we present some causes for the failure of approaches using simple co-occurrence measure for the determination of inter-term dependencies.

This paper has the following structure: In section 2 an assortment of simple, common co-occurrence measures and a set of linguistic phenomenons are presented. In section 3 the co-occurrence measures are evaluated by some samples and in section 4 a linguistic substantiation is given for the dissonances, which can be observed in section 3. Related work is presented in section 5 and finally a conclusion is drawn in section 6.

2. Definitions

Before going in details, we have to agree on a definition for the concepts ‘term’ and ‘word’. The concept *term* has it’s grounding in the common IF&IR literature and it stands for an atomic element of a document. In most IR&IF approaches documents are defined as a set (or sometimes a list) consisting of terms. Unlike terms, the concept of words is grounded in linguistics. In a linguistic sense a *word* is a unit of speech or writing that symbolizes or communicates a meaning. In most IF and IR systems terms are represented by exactly one word. For this reason we will use ‘term’ and ‘word’ synonymously.[1] ‘Term’ is used in this paper in the scope of IF and IR, while the ‘word’ is used in the scope of linguistics.

For the determination of inter-term dependencies using simple co-occurrence based measures a given document base has to be scanned for all combinations of joint appearance of two terms. We will define D as the set of all documents in the document base, and T as the set of all terms in the document base. $D_t \subseteq D$ is

[1] There exist in fact a minority of IF or IR systems which define a term as a set of words (e.g. some Language Models [SoCr1999]). These kinds of systems are not considered here.

the set of all documents containing the term $t \in T$. The dependency between two terms $a, b \in T$ is determined by the similarity sim(a, b) $\in$ [0…1] between these terms. Similarities near zero stand for independent terms (terms, which have nothing in common, like 'car' and 'mouse'), while similarities near one indicate a high term-dependency (for terms with a similar sense like 'car' and 'automobile'). The most frequently used measure for co-occurrence based term-similarity is the *Jaccard*-measure:

$$\mathrm{sim}(a,b) = \frac{\#D_{a\wedge b}}{\#D_a + \#D_b - \#D_{a\wedge b}}$$

with $D_{a\wedge b} = D_a \cap D_b$

Other simple measures are for example the *Dice*-measure

$$\mathrm{sim}(a,b) = \frac{2\cdot\#D_{a\wedge b}}{\#D_a + \#D_b}$$

and the *Cosine*-measure [MaSc1999] [Ferb2000].

$$\mathrm{sim}(a,b) = \frac{\#D_{a\wedge b}}{\sqrt{\#D_a \cdot \#D_b}}$$

Because IF and IR process natural language documents it is a good idea to validate the results given by the above statistical measures with our expectations from the linguistic point of view. To do this, we first have to figure out, which kind of linguistic dependencies between terms exist. Second we have to figure out which value the similarity sim(a, b) between two terms should have in an ideal case. Among others, the following linguistic dependencies (or phenomenon's) can be observed in natural languages like English or German:

Inflection is the change of word form according to a grammatical function. For example 'house' and 'houses' as well as 'fast', 'faster' and 'fastest' are different word forms of the same word. Compared to other linguistic phenomenon's inflection changes the meaning of words very little. For these reason the similarity of two words, which are word forms of the same basic form should be very high (near value one).

Synonymy denotes different words having the same meaning in some context. For example 'car' and 'automobile' have in most contexts the same meaning and can be substituted by each other. This causes that the similarity between two synonymous words should be very high (near value one).

Hyponymy is the semantic relation of subordination (is-a relationship). For example 'plant' is the subordinate of 'tree' or 'tulip'. Subordinate words are usually dependent to their superordinate words. This denotes, that the similarity between a subordinate and its superordinate should be relatively high. (Ideally, the similarity should depend on the number of intermediate subordination steps between both words.)

Meronymy is also known as the part-of relationship between words like 'wheel' and 'car'. Analogue to the case of Hyponymy the similarity between two words having the part-of relationship should be relatively high (and ideally depend on the number of intermediate steps between both words).

Composition is a linguistic phenomenon, which is seldom found in English, but it is common in some other languages for example like German. Composition describes the fact, that one word is composed of two or more other words. An English example for this phenomenon is 'submarine' which is composed of 'sub' and 'marine' or 'mastermind' which is composed of 'master' and 'mind'. In English composed words are seldom and usually the meaning of a composed word cannot be derived directly from its components. In the German language compositions are differently handled: A strict rule defines the semantics of composed words in that way, that a word 'xy' is always defined as an subordinate (hyponymy) of 'y' which is specialized by feature 'x'. For example the English translation of 'Gartenzwerg' is 'garden gnome', while 'Garten' means 'garden' and 'Zwerg' means 'dwarf' or 'gnome'. Because of this strict rule of interpretation of composed words, these kinds of words are very often used in the German language (and even new ad hoc compositions of words are allowed), which usually causes some trouble for IF and IR. The case of composition shows that the similarity of terms expected from the linguistic point of view is not only depending on linguistic phenomenon's but also on the language, which is processed. For the English language the similarity of a composed word to its components is very low or even null. For the German language the similarity of the composed word to its component should be relatively high (refer to argumentation for hyponymy).

Word group is the phenomenon of special groups of words, which have a different meaning to what someone would expect when just looking at the individual words. For example the word group 'New York' refers to a known American city, while 'new' means something that started its existence a short time ago and 'York' means a city in Britain. So the verbatim meaning of 'New York' should be the city in Britain which is called York, but which has been rebuild (for some reason) a short time ago. Another example is 'amber nectar', which stands for 'lager' or 'on carey street', which stands for 'being bankrupt'. Especially names of people, places and other things are often consisting of word groups. From the definition of word groups we derive, that the similarity of words, which are components of a word group, should usually have a value near zero.

In cases where no linguistic phenomenon between two words can be observed, the similarity of these words should have the value zero.

3. Samples

In this section the free encyclopedia Wikipedia [Wiki1] is used as document base to validate, if the above-presented statistical co-occurrence-based measures meet our expectations on similarities between pairs of terms from the linguistic point of view. The following tables base on a database snapshot [Wiki2] taken at the 14th of July 2004. Table 1 shows the co-occurrences of some terms of the English Wikipedia, which consisted of 745,546 individual documents at the time the snapshot has been taken. Some terms and their co-occurrences of the German Wikipedia are presented in Table 2. The German Wikipedia had 236,235 docu-

a	b	$\#D_a$	$\#D_b$	$\#D_{a \wedge b}$	Jaccard	Dice	Cosine	Phenomenon	exp. Sim.
New	York	89.488	23.717	22.283	0,245	0,394	0,484	word group	very low
Albert	Einstein	4.685	1.120	623	0,120	0,215	0,272	word group	very low
amber	nectar	480	287	4	0,005	0,010	0,011	word group	very low
Daniel	Winter	7.646	5.163	246	0,020	0,038	0,039	word group	very low
Bill	Gates	8.385	1.572	369	0,038	0,074	0,102	word group	very low
car	tree	6.889	6.887	298	0,022	0,043	0,043	none	null
house	red	20.530	16.764	2.063	0,059	0,111	0,111	none	null
sun	snail	6.717	266	28	0,004	0,008	0,021	none	null
office	mars	13.017	2.057	182	0,012	0,024	0,035	none	null
office	mouse	13.017	1.592	151	0,010	0,021	0,033	none	null
submarine	sub	2.544	8.493	186	0,017	0,034	0,040	composition	low
submarine	marine	2.544	3.814	261	0,043	0,082	0,084	composition	low
mastermind	master	186	6.633	28	0,004	0,008	0,025	composition	low
mastermind	mind	186	10.379	35	0,003	0,007	0,025	composition	low
wheel	car	2.139	6.889	517	0,061	0,115	0,135	meronymy	high
body	leg	14.850	1.447	370	0,023	0,045	0,080	meronymy	high
hand	finger	14.254	1.267	461	0,031	0,059	0,108	meronymy	high
computer	memory	17.077	5.876	1.658	0,078	0,144	0,166	meronymy	high
computer	CPU	17.077	933	636	0,037	0,071	0,159	meronymy	high
plant	tree	6.155	6.887	968	0,080	0,148	0,149	hyponymy	high
plant	tulip	6.155	134	32	0,005	0,010	0,035	hyponymy	high
planet	earth	5.067	12.855	1.994	0,125	0,223	0,247	hyponymy	high
planet	mars	5.067	2.057	911	0,147	0,256	0,282	hyponymy	high
OS	windows	2.105	3.596	632	0,125	0,222	0,230	hyponymy	high
OS	linux	2.105	2.309	490	0,125	0,222	0,222	hyponymy	high
car	automobile	6.889	2.641	1.187	0,142	0,249	0,278	synonymy	very high
hope	esperance	16.887	60	8	0,000	0,001	0,008	synonymy	very high
pretty	beautiful	7.022	3.868	365	0,035	0,067	0,070	synonymy	very high
petit	small	594	43.605	245	0,006	0,011	0,048	synonymy	very high
goal	target	4.740	4.799	376	0,041	0,079	0,079	synonymy	very high
house	houses	20.530	4.184	1.663	0,072	0,135	0,179	inflection	very high
mouse	mice	1.592	521	206	0,108	0,195	0,226	inflection	very high
fast	faster	5.485	2.652	549	0,072	0,135	0,144	inflection	very high
fast	fastest	5.485	907	182	0,029	0,057	0,082	inflection	very high
car	cars	6.889	2.770	1.456	0,177	0,301	0,333	inflection	very high

Table 1: Co-occurrences of some terms in the English Wikipedia.

ments at the time the snapshot has been taken. For processing, all letters of the documents and queries has been transformed into lower-case letters.

On examining table 1 the relatively low values of all similarities is striking. The highest similarity exists between the terms ‘New’ and ‘York’. Depending on the measure used, the similarity value of this term-pair is 0.245 (Jaccard), 0.394 (Dice) or 0.484 (Cosine). This is surprising because from the linguistic point of view the similarity for ‘New’ and ‘York’ (word group) should not be very high. Rather it should be at least significantly lower then the similarities of term-pairs, which are linked by linguistic phenomenons like synonymy or inflection. In fact it is striking, that a lot of term-pairs (e.g. ‘body’ and ‘leg’ (meronymy), ‘hand’ and ‘finger’ (meronymy), ‘computer’ and ‘CPU’ (meronymy), ‘plant’ and ‘tulip’ (hyponymy), ‘petit’ and ‘small’ (synonymy) and ‘fast’ and ‘fastest’ (inflection)) which are expected to have significantly high similarities from the linguistic point of view have a lower similarity regarding the co-occurrence measures than

a	b	$\#D_a$	$\#D_b$	$\#D_{a \wedge b}$	Jaccard	Dice	Cosine	Phenomenon	exp. Sim.
New	York	4908	3158	2936	0,572	0,728	0,746	word group	very low
Albert	Einstein	1851	504	329	0,162	0,279	0,341	word group	very low
Hans	Müller	4491	1395	451	0,083	0,153	0,180	word group	very low
Hasso	Plattner	21	18	6	0,182	0,308	0,309	word group	very low
Bill	Gates	781	134	64	0,075	0,140	0,198	word group	very low
Auto	Baum	1288	1411	67	0,025	0,050	0,050	none	null
Haus	rot	3138	1961	153	0,031	0,060	0,062	none	null
Sonne	Schnecke	1746	92	6	0,003	0,007	0,015	none	null
Büro	Mars	371	613	14	0,014	0,028	0,029	none	null
Büro	Maus	371	494	7	0,008	0,016	0,016	none	null
Gartenzwerg	Zwerg	20	220	3	0,013	0,025	0,045	composition	high
Gartenzwerg	Garten	20	906	0	0,000	0,000	0,000	composition	high
Hausmeister	Meister	64	1541	5	0,003	0,006	0,016	composition	high
Hausmeister	Haus	64	3138	16	0,005	0,010	0,036	composition	high
Reifen	Auto	323	1288	28	0,018	0,035	0,043	meronymy	high
Körper	Fuß	2597	1086	64	0,018	0,035	0,038	meronymy	high
Hand	Finger	2834	532	160	0,050	0,095	0,130	meronymy	high
Computer	Speicher	2858	491	178	0,056	0,106	0,150	meronymy	high
Computer	CPU	2858	293	157	0,052	0,100	0,172	meronymy	high
Pflanze	Baum	1157	1411	144	0,059	0,112	0,113	hyponymy	high
Pflanze	Tulpe	1157	29	2	0,002	0,003	0,011	hyponymy	high
Planet	Erde	1093	3130	415	0,109	0,197	0,224	hyponymy	high
Planet	Mars	1093	613	288	0,203	0,338	0,352	hyponymy	high
Betriebssystem	Windows	935	3445	362	0,090	0,165	0,202	hyponymy	high
Betriebssystem	Linux	935	1201	338	0,188	0,316	0,319	hyponymy	high
Auto	Automobil	1288	387	122	0,079	0,146	0,173	synonymy	very high
Computer	Rechner	2858	741	277	0,083	0,154	0,190	synonymy	very high
senkrecht	vertikal	493	168	19	0,030	0,057	0,066	synonymy	very high
Fahrstuhl	Lift	33	55	2	0,023	0,045	0,047	synonymy	very high
Orange	Apfelsine	561	32	16	0,028	0,054	0,119	synonymy	very high
Haus	Häuser	3138	795	231	0,062	0,117	0,146	inflection	very high
Maus	Mäuse	494	140	49	0,084	0,155	0,186	inflection	very high
schnell	schneller	4369	1255	331	0,063	0,118	0,141	inflection	very high
schnell	schnellsten	4369	158	33	0,007	0,015	0,040	inflection	very high
Auto	Autos	1288	430	141	0,089	0,164	0,189	inflection	very high

Table 2: Co-occurrences of some terms in the German Wikipedia.

some term-pairs which are expected to have low similarities (e.g. ‘Albert’ and ‘Einstein’ (word group), ‘house’ and ‘red’ (no linguistic phenomenon exists)).

On examining table 2 we see a similar situation for the German language. In this table, the highest similarity is also found between the terms ‘New’ and ‘York’. Depending on the used measure, the similarity value of this term-pair is 0.572 (Jaccard), 0.728 (Dice) or 0.746 (Cosine). The very high value of the similarity can be explained by the fact, that ‘New’ and ‘York’ are not domestic words of the German language. A closer look at table 2 shows, that e.g. ‘Haus’ and ‘rot’ (no linguistic phenomenon) have a higher similarity (although the similarity is relatively low) than for example ‘Gartenzwerg’ and ‘Zwerg’ (composition), ‘Reifen’ and ‘Auto’ (meronymy) or ‘Pflanze’ and ‘Tulpe’

(hyponymy) or 'Fahrstuhl' and 'Lift' (synonymy) or 'schnell' and 'schnellsten' (inflection) which are expected to have relatively high similarities.[2] It is worth to mention that in contrast to the English language the co-occurrences for compositions for the German language do not match expected similarities, which reduces the quality of co-occurrence measures for the German language in comparison to the measures for the English language. Hence, it can be recorded that the quality of co-occurrence measures is language dependent.

Regarding to the samples of these both tables we have to accept the fact, that in most samples all three co-occurrence measures do not meet the similarities expected from the linguistic point of view. And what is even worse, there some term-pairs existing, which are expected to have significantly lower similarities (from the linguistic point of view) than some others, but which have in fact higher similarities than the others when simple co-occurrence measures are used. Thus it is impossible to meet the linguistic motivated claims of term similarities by using a monotonic transformation of the presented co-occurrence measures, which makes the co-occurrence based similarities useless for the determination of inter-term dependencies.

4. Linguistic Substantiation

It is possible to set up a position, that the samples presented in the tables represent only some rare and special cases, so that simple co-occurrence measures will not suffer from wrong similarities when the document base is large enough or when other term-pairs are considered. In this section we will show step by step how unlikely it is, that this position holds.

Beginning with inflection, it can be noted in practical experience that especially in short documents only one or a few inflected forms of a word are used. For instance, if you write a document about a special instance of a thing ('house', 'car', 'computer', etc.), you will usually not need the plural form very often, because you are writing about *one* special thing. This results in the fact that plural and singular word forms will tend to have lower co-occurrence based similarities than it is expected from the linguistic point of view. This argumentation can be expanded to most other forms of flexions. Consequently this coherence is the reason why most co-occurrence based approaches benefit from stemming.

Regarding synonymy, the practical experience shows that the co-occurrence of two synonymous words depends on the kind of literature they are embedded. To avoid iterations the usage of synonyms is quite common in narrative literature. For academic literature the opposite is true, unnecessary synonyms are usually avoided to prevent misunderstanding[3]. Especially for seldom words the

[2] A free German-English dictionary is available at the following web page: http://dict.leo.org/?lang=en

[3] In fact, synonyms are one source for misunderstanding between different scientific areas. Often the same things have different technical terms in different areas.

probability of the co-occurrence of their synonym within the same text is low, because there is only a marginal probability for the iteration of the same word or its synonym. For this reasons simple co-occurrence based similarities for seldom synonyms must be too low in general.

In usual documents it is not common to enumerate all subordinates of a thing/word (hyponymy) or all parts of something (meronymy). Only special documents, which focus on these kinds of enumerations like construction manuals or ontology focused documents enumerate subordinates or parts of things in a detailed manner. This results in the fact, that words related by hyponymy or meronymy will suffer from too low similarities when co-occurrence based measures are used.

While compositions in the English are well represented by simple co-occurrence based approaches, their similarity is too low in general for the German language. This can be ascribed to the fact, that compositions in German are always hyponyms (refer to section 2). So the above explanation for hyponymy holds up for this case.

As simple co-occurrence based approaches tend to underestimate the similarities between term pairs in the above-presented cases, these approaches tend to overestimate the similarities for word groups. This overestimation worsens, the more popular a word group is. Popular word groups result in a high combined occurrence of the word group components in relation to the occurrence of only one alone component. This leads to the mentioned overestimation of word groups.

As the bottom line we have to accept, that simple co-occurrence based approaches are not suitable for the determination of similarities of term-pairs.

5. Related Work

In [OgMK1991] OGAWA, MORITA and KOBAYASHI present a fuzzy retrieval system. One feature of this system is the *keyword-connection-matrix*, which is used to represent similarities of term-pairs. To optimize their retrieval system the authors implemented an adaptive approach to modify the entries in the keyword-connection-matrix using a gradient decent. A set of predefined queries and a list of adequate documents has been used as training data for the adaptive approach. Prior to training, the keyword-connection-matrix has been initialized by similarities provided by the Jaccard co-occurrence measure.

While optimizing their retrieval system the authors noticed, that the retrieval quality of their system rose significantly during the training. Further, when looking at their documentation, a major change of the values in the keyword-connection-matrix can be observed. This fact holds up the thesis presented in this paper. The major change of the values in the keyword-connection-matrix attest that the co-occurrence initialized keyword-connection-matrix was far away from the optimum. This is an evidence for the uselessness of co-occurrence based measures for term-pair similarities in information filtering und retrieval.

6. Summary, Conclusion and Future Work

Within this paper, we give a motivation for the use of term-pair similarities for the improvement of information filtering and retrieval methods. We presented some simple, but often used co-occurrence based measures to gain similarity values from a document base. In the main part of the paper we show by an example and by linguistic substantiation that simple co-occurrence based measures for derivation of paired term similarities provide bogus similarity values for the English and German language. These approaches tend to underestimate similarities for terms linked by inflection, synonymy, hyponymy and meronymy, while the same approaches tend to overestimate similarities for word groups. Further a language dependency of the quality of co-occurrence measures has been determined.

From these results we have to conclude, that simple co-occurrence based approaches should not be used for the estimation of term-pair similarities for information filtering and retrieval methods, because an improvement of these methods by co-occurrence based measures is implausible.

This paper discusses only simple co-occurrence based approaches, but there are more sophisticated approaches existing, which try to deal with those various linguistic phenomenons by implicitly or explicitly deriving term similarities from the arrangement of terms in relation to documents. For example the Latent Semantic Index model [FDDL1988] and the Language Model [SoCr1999] embed such approaches. From the scientific point of view it will be interesting to analyze if those approaches are able to meet the term-similarity expectations from the linguistic point of view or if they also suffer from the same problems like the simple co-occurrence based approaches.

7. References

1. [BaRi1999] R. Baeza-Yates, B. Ribeiro-Neto: Modern Information Retrieval. Addision Wesley Publishing Company, 1999.
2. [BeCr1992] N. J. Belkin, W. B. Croft: Information Filtering and Information Retrieval: Two Sides of the Same Coin? In: Communications of the ACM, 35(12), 1992, pp. 29-38.
3. [Belew1989] R. Belew: Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In: Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1989, pp. 11-20.
4. [BeKu2003] J. Becker, D. Kuropka: Topic-based Vector Space Model. In: Proceedings of the 6th International Conference on Business Information Systems, 2003, pp. 7-12.
5. [FDDL1988] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, et al.: Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1988, pp. 465-480.

6. [Ferb2000] R. Ferber: Data Mining und Information Retrieval. 2000. http://information-retrieval.de/dm-ir
7. [Kuro2004] D. Kuropka: Modelle zur Repräsentation natürlichsprachlicher Dokumente – Information-Filtering und -Retrieval mit relationalen Datenbanken, Logos, Berlin, 2004.
8. [Kwok1989] K. L. Kwok: A neuronal network for probabilistic information retrieval. In: Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1989, pp. 21-30.
9. [Mand2001] T. Mandl: Tolerantes Information Retrieval: neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. UVK-Verlagsgesellschaft, Konstanz, 2001.
10. [MaSc1999] C. D. Manning, H. Schütze: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, 1999.
11. [RoJo1976] S. E. Robertson, K. S. Jones: Relevance weighting of search terms. In: Journal of the American Society for Information Sciences, 27(3), 1976, pp. 129-146.
12. [Salt1968] G. Salton: Automatic Information Organization and Retrieval. McGraw-Hill, New York, 1968.
13. [SoCr1999] F. Song, W. B. Croft: A General Language Model for Information Retrieval. In: Proceedings on the 8th International Conference on Information and Knowledge Management (CIKM'99), 1999, pp. 316-321.
14. [OgMK1991] Y. Ogawa, T. Morita, K. Kobayashi: A fuzzy document retrieval system using the keyword connection matrix and a learning method. In: Fuzzy Sets and Systems (39), 1991, pp. 163-179.
15. [WiHi1991] R. Wiklinson, P. Hingston: Using the cosine measure in neuronal network for document retrieval. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1991, pp. 202-210.
16. [Wiki1] Wikipedia: The Free Encyclopedia. http://www.wikipedia.org
17. [Wiki2] Wikipedia database download. http://download.wikimedia.org
18. [WZRW1987] S. K. Wong, W. Ziarko, V. V. Raghaven, R. C. N. Wong: On Modeling of Information Retrieval Concepts in Vector Spaces. In: ACM Transactions on Database Systems, 12(2), 1987, pp. 299-321.