Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 249–256

ISBN 978-83-60810-51-4

A Study of Measures for Document Relatedness Evaluation

Evgeny Pyshkin
Saint-Petersburg State Polytechnical University
Department of Computer Systems and Software Engineering
21 Politekhnicheskaya st., St.-Petersburg, 194021
Russian Federation
Email: pyshkin@ftk.spbstu.ru

Vitaly Klyuev
University of Aizu
Division of Information Systems
Software Engineering Laboratory
Tsuruga Ikki-machi, Aizu-Wakamatsu City,
Fukushima, 965-8580, Japan
Email: vkluev@u-aizu.ac.jp

***Abstract*—In this review paper we classified and described measures and approaches for document relatedness evaluation. For the reviewed measures we pointed out the reasons of their construction and usage limitations. We concluded this research with a discourse on challenges of the day in estimating document appropriateness in the domain of information retrieval.**

***Keywords*—informational search; document relatedness; semantic similarity; semantic measures; information retrieval**

I. Introduction

IN TEXT information retrieval and data mining we evaluate retrieved documents so to find out their similarity and their relevance to users' demands while searching. It is true that the quality of searching process depends not only on the algorithms used by search engines, but also on quality of users' queries and on the assisting tools that help to expand or modify queries to reach search goals better. Measures used to evaluate relatedness are not only the fundamental part of searching algorithms but also a basic mathematical method to support decision making while considering document aboutness, document relevance, and document semantics in natural language processing. Rinaldi discriminates two main classes of information relevance: objective (system-based) and subjective (user-based) [1]. Objective relevance is mostly focused on measuring topic or concepts matching for the retrieved documents and the query, while subjective relevance refers to the user interpretation and relates to the concepts of aboutness and appropriateness.

Considering particularly linguistic aspects of a relevance concept, as far as 1996 Saracevic proposed five types of relevance: an algorithmic relevance between the query and retrieved documents, a topic-based relevance associated with the concept of aboutness, cognitive relevance for measuring documents from tuser's view, situational reference related to the intellectual interpretation of the search tasks, and affective relevance, focused on achieving the search goals [2]. It is clear that measures used to estimate documents serve not only one, but different relevance classes as a rule.

Since 1999 we know some comprehensive overviews of known methods and approaches appeared, including extensive works of Resnik, 1999 [3], Budanitsky and Hirst, 2005 [4], [5], and Rinaldi, 2009 [1]. Valuable extensions was added by Harrington in respect to the classification of approaches aimed to measure relatedness with introducing resource based methods and distributional methods [6]. Our study is aimed to classify principal approaches of measuring document relatedness with taking into the consideration some recent works in the domain.

This work is supported by the University of Aizu

II. Measuring Relatedness

In this section we review and analyze existing approaches, from the classic distance based similarity to the relatively complex algorithms which combine components related to different classes of measures.

A. Distance-based similarity and relatedness measuring

Let us begin with the terms or word relatedness evaluation based on concept semantic relatedness calculated by term relative positions in lexical taxomomies. General schema is the following [3], [5]:

$$rel(w_1, w_2) = max_{c_1 \in S(w_1), c_2 \in S(w_2)}[rel(c_1, c_2)]$$

$S(w_i)$ being the set of concepts in the taxonomy that are senses of w_i .

The approach to evaluate document relevance proposed by Leacock and Chodorow in [7] is based on the word similarity metric (LC-measure):

$$\mathrm{sim}_{LC}(w_1, w_2) = max[-\log \frac{length(w_1, w_2)}{2\mathrm{d}_{max}}] \quad ,$$

$length(w_1, w_2)$ being number of nodes in is-a hierarchy from w_1 to w_2 and d_{max} being maximum depth of the taxonomy.

Altintas et al. remarked that one problem of the Leacock and Chodorow's measure is that it is not able to differentiate concepts having the same shortest path from an input concept [8]. So they proposed to improve evaluation by taking into accounts not only the concept commonness but also their relative specificity:

$$Spec(c) = \frac{Depth(c)}{ClusterDepth(c)} \quad ,$$

$ClusterDepth(c)$ being the depth of the deepest node in the cluster (sub-tree containing respective concept node).

Resulting similarity measure (that we refer as ALT-measure) is based on two components, one – for length based measure (similar to the sim_{LC}), and other – for specificity based measure:

$$lengthFactor = \frac{length(c_1, c_2)}{2d_{max}}$$

$$specFactor = |Spec(c_1) - Spec(c_2)|$$

$$sim_A(c_1, c_2) = \frac{1}{1 + lengthFactor + specFactor}$$

Wu and Palmer also use the similarity measure based on is-a hierarchy, but their measure (referred here as WP-measure) computes the distance to the root node of the most specific concept $lcs(c_1, c_2)$ that intersects the path of two concepts (c_1, c_2) in is-a hierarchy [9]:

$$sim_{WP}(c_1, c_2) = 2 \cdot \frac{depth(lcs(c_1, c_2))}{length(c_1, lcs(c_1, c_2)) + length(c_2, lcs(c_1, c_2)) + 2 \cdot depth(lcs(c_1, c_2))},$$

where *depth* is the distance from the concept node to the root node, and *length* is the path length between concepts.

Hirst and St-Onge proposed to classify relations in accordance to their strength into *extra-strong*, *strong*, *medium-strong* and *weak*, together with help of some geometric representation of relationships [10]. For medium-strong relations they proposed to compute the taxonomic path length used to evaluate concept relatedness. To estimate the weight for the path between two nodes they count not only hops between nodes, but also how many changes of directions are for the path:

$$rel_{HS}(c_1, c_2) = C - length(c_1, c_2) - k \cdot turns(c_1, c_2),$$

where *C* and *k* are empirically defined constants (Hirst and St-Onge uses *C=8* and *k=1* respectively, $turns(c_1, c_2)$ is the number of times the path between c_1 to c_2 changes direction).

Hirst and St-Onge also defined patterns for allowable paths which should be evaluated. Some rules describing constructing paths expressing reasonable relations are foundations for these patterns:

- No other relation may precede the upward link;
- At most one change of direction is allowed;
- It is permitted to use a horizontal link to make a transition from an upward to a downward direction.

Let's note that the classification proposed by Hirst and St-Onge has evident analogy with object hierarchies in object-oriented languages and especially with up-casting and down-casting concepts.

B. Approaches based on information theory

Resnik remarked that using uniform edge-based or node-based distance counting may be unreliable for the reason that some sub-taxonomies may be denser than others [3]. To overcome the unreliability of direct edge counting he proposed to augment the taxonomy with some probabilistic function $p: C \rightarrow [0,1]$, such that for any concept $c \in C$, $p(c)$ is probability of encountering an instance of concept *c*. As a result, the following measure based on the information theory argumentation has been proposed:

$$sim_{RI}(c_1, c_2) = max_{c \in S(c_1, c_2)}[-\log(p(c))],$$

$S(c_1, c_2)$ being the set of concepts that subsume both c_1 and c_2.

Then for word similarity the following measure may be used:

$$wsim_{RI}(w_1, w_2) = max_{(c_1, c_2)}[sim_{RI}(c_1, c_2)],$$

where c_1 ranges over $S(c_1)$ which is set of concepts subsuming c_1, and c_2 ranges over $S(c_2)$ which is set of concepts subsuming c_2.

Instead of using negative values produced by the measure using the information content, the probability concept may be directly used:

$$sim_{RP}(c_1, c_2) = max_{c \in S(c_1, c_2)}[1 - p(c)]$$

Respectively:

$$wsim_{RP}(w_1, w_2) = max_{(c_1, c_2)}[sim_{RP}(c_1, c_2)]$$

Function $p(c)$ is a sort of implementation depending function. For the WordNet taxonomy model frequencies of concepts in the taxonomy are being used to compute $p(c)$ values:

$$freq(c) = \sum_{n \in words(c)} count(n),$$

where $words(c)$ is set of words subsumed by concept *c*. Concept probabilities may be computed as relative frequency:

$$\hat{p}(c) = \frac{freq(c)}{N},$$

where *N* is a total number of nouns observed (excluding those non subsumed by any WordNet class).

Budanitsky mentioned that for the reason that Resnik's approach attempts to deal with problem of varying link distances by generally downplaying the network edges in the determination of the degree of semantic proximity, using rather selectively the taxonomy structure, some semantic relations may be indistinguishable, especially in case when we analyze two pairs of concepts having the same most-specific subsumer (the fact that the number of edges is different for different branches of the taxonomy tree is not taken into the consideration, see [3], [5] for details).

C. Hybrid approaches for similarity and relatedness estimation

Resnik noticed that it is important to distinct between semantic relatedness and semantic similarity (which represents the special case of relatedness and deals mostly with cognitive synonyms). Jiang and Conrath tried to combine edge- and node-based methods by using edge-counting as a dominant strategy, and statistical information as a corrective factor [11].

Rinaldi introduces Dynamic Semantic Network (DCN) built dynamically in process of semantic analysis on the base of some dictionary (e.g. WordNet) [1]. Similar to Jiang and Conrath, Rinaldi's approach also combines statistical information with taxonomy data. Semantic network is being constructed in accordance to the definition given by Lee et al. in [12], where semantic network is considered as a graph consisting of nodes representing terms and concepts and edges representing semantic relations.

Harrington introduced an approach using semantic networks created by general purpose NLP tools with subsequent computation of spreading activations in order to evaluate the relatedness strength of analyzed terms [6]. What's the promising about this solution is that it is obtained as a side effect of one general purpose tool for information collection which was not specially constructed to determine relatedness. It seems that it's human to proceed with semantic

analysis in this way, from general understanding of some matter to its concrete applications.

Varelas et al. introduced single ontology method called semantic similarity retrieval model (SSRM) which combines statistical information of terms usage and information about terms position in some taxonomy [13]. The similarity metric they used, is based on the TF-IDF model together with vector space model (VSM) [14], [15]. According to TF-IDF model, the weight of the term in the collection of documents is computed as follows:

$$w_i(d)=\mathrm{tf}_i\cdot\mathrm{idf}_i \quad ,$$

where t_i is the i-th term in the document d, tf_i is the frequency of t_i in the document (i.e. number of times the term appears in the document), idf_i is the inverse frequency of t_i in the collection of documents (i.e. number of documents where this term is present).

Adapting the vector space model, the following metric may serve the similarity between two documents:

$$\mathrm{sim}_{VSM}(d_1,d_2)=\frac{\sum_{i=1}^{NT} w_i(d_1)\cdot w_i(d_2)}{\sqrt{\sum_{i=1}^{NT} w_i(d_1)^2}\cdot\sqrt{\sum_{i=1}^{NT} w_i(d_2)^2}} \quad ,$$

where NT is the dimensionality of vector space (number of different terms).

Then the authors proposed correction of term weighting on the base of its connections with other semantically related terms (they used WordNet ontology, but the idea is general):

$$w_i^{(SSRM)}(d)=w_i(d)+\sum_{\mathrm{sim}_{VSM}(i,j)\ge T}^{j\ne i} w_i(d)\cdot\mathrm{sim}_{VSM}(i,j) \quad ,$$

where T is the user defined threshold (authors used $T=8$).

After re-weighting query terms, the query itself is being expanded with using synonym terms, together with hyponyms and hypernyms which are semantically similar within the similarity limited by the user threshold. Then, each term of the new query is assigned a new weight as follows:

$$w_{(i,\exp)}^{(SSRM)}(q)=\sum_{\mathrm{sim}_{VSM}(i,j)\ge T}^{j\ne i}\frac{1}{N_{j,hyp}} w_j(q)\cdot\mathrm{sim}_{VSM}(i,j) \quad ,$$

for expansion terms;

$$w_{(i,\exp)}^{(SSRM)}(q)=w_i^{(SSRM)}(q)+\sum_{\mathrm{sim}_{VSM}(i,j)\ge T}^{j\ne i}\frac{1}{N_{j,hyp}} w_j(q)\cdot\mathrm{sim}_{VSM}(i,j) \quad ,$$

for initial terms. Here $N_{j,hyp}$ is the number of hyponyms for each expanded j-th term. For hypernyms $N_{j,hyp}=1$.

After query expansion the similarity between a reweighted query $q_{\exp}$ and a document d is defined as follows:

$$\mathrm{sim}_{SSRM}(q_{\exp},d)=\frac{\sum_i\sum_j w_{(i,\exp)}^{(SSRM)}(q_{\exp})w_j(d)\cdot\mathrm{sim}_{VSM}(i,j)}{\sum_i\sum_j w_{(i,\exp)}^{(SSRM)}(q_{\exp})w_j(d)}$$

where i and j spans for terms of the query and of the document respectively. It should to be noticed that unlike to the query terms, document terms are neither expanded not re-weighted.

Let us note that for those approaches dealing directly with taxonomy distance based information that have been tested on the WS-353 test collection, the Spearman ρ rank correlation coefficient between the relatedness ranking of pairs by human judges and that by the tested measure ranges between 0.13 and 0.31 [16], so there is space to further improvement.

D. Dictionary-based approaches

Kozima and Ito constructed context-sensitive dynamic measurement of the word distance on the base of semantic space adaptive scaling [17].

This is a dictionary-based algorithmic approach. The principal idea is that the word distance changes in different contexts. For example, in one context "engine" may be near the "car" (considering the context of car composition); while in other context (e.g. considering the context of means of transportation) it may be much more distant in comparison with "bus" or "railway".

In this approach every dictionary word is being mapped onto NDV-dimensional vector, called P-vector, which is the representation of the word in basis of defining vocabulary, NDV being the number of words in the defining vocabulary (all definitions in the dictionary are written using the words in the defining vocabulary or words defined elsewhere in the dictionary). So, for given word w the vector $P(w)$ represents the meaning of the word in its relationship to other words. Semantic distance between two words may be computed as geometric distance between respective P-vectors. But this measure is not context-sensitive. To decrease complexity and to eliminate the noise, every P-vector is being mapped to the Q-vector with using smaller amount of dimensions. The general algorithm described is as follows:

1. Compute principal components $X_1, X_2, \dots, X_{NDV}$ – each of which is NDV-dimensional vector – under the following conditions:
   a) For any $X_i, i=1\div NDV$ its norm $|X_i|=1$.
   b) For any pair $\{X_i, X_j\}, i\ne j$ their inner product is equal to 0.
   c) The variance v_i of P-vectors projected onto X_i is not smaller than any $v_j, j>i$ (it means that X_1 is the first principal component with the largest variance, X_2 is the second principal component with the second-largest variance, and so on.
2. Select first MDV principal components $X_1, X_2, \dots, X_{MDV}$, where $MDV \ll NDV$.
3. Map each NDV-dimensional P-vector onto MDV-dimensional Q-vector in basis $\{X_1, X_2, \dots, X_{MDV}\}$.

Authors proposed to scale weights for each dimension up or down so to make word forming a cluster in the semantic space. This cluster refers to the semantic context. Then the distance between two words under the context C is defined as follows:

$$d(w_1,w_2\mid C)=\sqrt{\sum_{i=1}^{MDV}(f_i q_{i,1}-f_i q_{i,2})^2} \quad ,$$

where $q_{i,k}$ is the i-th dimension of the Q-vector for the k-th word, $f_i\in[0,1]$ is the scaling factor of the i-th dimension defined as follows:

$f_i=1-r_i$ for $r_i\le 1$, otherwise $f_i=0$;

$r_i=\frac{SD_i(C)}{SD_i(V)}$, where $SD_i(C)$ being the standard deviation of the i-th component of $w_1, w_2, \dots, w_{NDV}$;

$SD_i(V)$ is that of words in the whole defining vocabulary.

Another interesting research which uses the vector representation of the concepts in the semantic network was introduced by Baziz et al. [18]. At the first step, the single- and multi-word concepts belonging to some ontology (e.g. WordNet) are extracted from given document. The extracted concepts are weighted on the base of concept frequency factor. The concept frequency is computed as an adopted TF-IDF construct by taking into the consideration the number of occurrences of the concept itself in the documents and weighted sum of occurrences of all sub-concepts:

$$cf(c_i)=count(c_i)+\sum_{sc\in \mathrm{sub}(c_i)} \frac{nw(sc)}{nw(c_i)} count(sc) \quad ,$$

where $cf(c_i)$ is the concept frequency for the concept c_i, $count(c)$ is the number of occurrences of the concept in the document, $\mathrm{sub}(c_i)$ is the set of all sub-concepts for c_i, and $nw(c)$ is the number of words in concept description.

For example, for the concept "wheeled motor vehicle" composed of three words the concept frequency is as follows:

$$\begin{aligned} cf(\text{wheeled motor vehicle}) &= count(\text{wheeled motor vehicle}) \\ &+\frac{2}{3} count(\text{wheeled vehicle}) \\ &+\frac{2}{3} count(\text{motor vehicle}) \\ &+\frac{1}{3} count(\text{wheeled}) \\ &+\frac{1}{3} count(\text{motor}) \\ &+\frac{1}{3} count(\text{vehicle}) \end{aligned}$$

Then the weight of the concept c_i is considered as global frequency of the concept in the document d_i:

$$w(c_i,d_j)=cf(c_i)\cdot \ln\left(\frac{ND}{idf_i}\right) \quad ,$$

where *ND* is the total number of documents, idf_i is the number of document containing the concept.

This measure is used to rank concepts for selection of the concepts to include into the document semantic core.

At the second step the concept senses and their semantic relatedness are to be identified. Suppose the set of concepts selected at the first stage:

$$D_C=\{C_i\}, i=1\div N_C \quad ,$$

where N_c is the number of selected concepts.

For every concept its senses have been discovered from the ontology (e.g. from WordNet synsets):

$$S^{(i)}=\{S_j^{(i)}\}, j=1\div N_S^{(i)}$$

where $N_S^{(i)}$ being the number of senses discovered for the concept C_i.

Suppose we have N_R types of relations in the ontology, so:

$$R=\{R_i\}, i=1\div N_R$$

The semantic relatedness between two concept senses is estimated as adopted Lesk intersection (number of common words which is squared in case of successive words found in strings representing the information returned for all relations from ontology, when applied to the given concept senses, see also [19]):

$$\mathrm{rel}_{BAZ}(S_{(j_s)}^{(k)}, S_{(j_p)}^{(l)})=\sum_{(i,j)\in\{1,\ldots,N_R\}} R_i(S_{j_s}^{(k)})\cap R_j(S_{j_p}^{(k)})$$

Then the semantic network may be defines as weighted graph containing nodes representing concept senses:

$$SN(j)=\{S_{j_i}^{(i)}\}, i=1\div N_C, j_i=1\div N_S^{(i)} \quad ,$$

where $SN(j)=\{S_{j_i}^{(i)}\}, i=1\div N_C, j_i=1\div N_S^{(i)}$ being the j-th configuration of concept senses from D_C, j_i is the sense index ranging from 1 to all possible senses for respective concept C_i.

The edges represent semantic relatedness between concepts and marked by rel_{BAZ} values. Every document may have many possible semantic networks, so to compose the best one, the concept senses are being chosen maximizing the total relatedness score:

$$score(S_k^{(i)})=\sum_{l\in[1..N_C],\, l\neq i,\, k\in[1..N_S^{(l)}]} \mathrm{rel}_{BAZ}(S_k^{(i)}, S_j^{(l)})$$

$$best(C_i)=max_{k=1..N_C} score(S_k^{(i)})$$

So the final semantic core is a graph where nodes correspond to selected concepts and marked by the best score values.

The approaches presented in this section, seems to be very useful for such problems as concepts extraction and disambiguation, query expansion by choosing relevant terms and concepts, measuring semantic distance between concepts, and identifying concepts describing document content better.

E. Wikipedia-based approaches

In recent years Wikipedia rapidly became one of the most important sources of knowledge. Being the sort of collaborative knowledge base, Wikipedia is currently the largest repository of world knowledge. The number of articles in the English version is at least 15 times higher than in Encyclopedia Britannica. Despite the fact that the accuracy and the quality of Wikipedia still remain subject of many discussions, there is not controversial to use Wikipedia as a source of lexical semantic knowledge [12], [20], [21].

One obvious idea is to use Wikipedia as a lexical resource (e.g. instead of WordNet or similar lexical bases) and to proceed with the path based measurements described in the above sections. This idea has been implemented by Strube and Ponzetto and shown better results comparing to similar implementations using WordNet [22]. Herrington mentioned that one explanation of this is the nature of Wikipedia, where articles have links to other articles which are not simply similar but evidently related [6].

In [20], [23], [24] authors used vector representation of the analyzed document in accordance to TF-IDF model and estimate similarity between documents by regular cosine metric for corresponding vectors that is similar to $\mathrm{sim}_{VSM}(d_1,d_2)$ described earlier. This method is known as explicit semantic analysis (ESA). The semantic

interpretation vector S for some documents D is defined as follows:

$$S=\{s_i, i=1 \div N_C\}, s_i=\sum_{(w_j \in D)} tf_j^{(ESA)} \cdot idf_i^{(ESA)}$$

where N_C is the total number of Wikipedia concepts; w_j is the j-th word in the document D; $tf_j^{(ESA)}$ is the weight for word w_j; $idf_i^{(ESA)}$ is the inverted index entry for word w_j (the strength of association of word w_j with Wikipedia concept $c_i, i=1 \div N_C$).

Instead of counting terms, Milne and Witten proposed to count links [25]:

$$lw(s \to t)=\log(\frac{|W|}{|T|}), s \in T \quad .$$

where s and t are the source and target articles; $lw(s \to t)$ is the weight of the link $s \to t$; T is the set of articles having links to t; W is the set of all Wikipedia articles ($lw(s \to t)=0, s \notin T$).

Link weights are used to generate vectors describing documents to compare in the same manner as in the ESA method.

The second measure used by Milne and Witten is a link-based adaptation of the normalized Google similarity distance metric:

$$sr(a,b)=\frac{\log(max(|A|,|B|))-\log(|A \cap B|)}{(\log(|W|)-\log(min(|A|,|B|)))} \quad ,$$

where a and b are two articles of interest, A and B are the sets of all articles that link to a and b respectively, W is the set of all Wikipedia articles.

The original measure for normalized Google distance between terms t_1 and t_2 introduced in [26] uses information about terms distribution, rather than the amount of links and is defines as follows:

$$dist_{Google}(t_1,t_2)=\frac{max(\log(|T_1|),\log(|T_2|))-\log(|T_1 \cap T_2|)}{\log G-min(\log(|T_1|,|T_2|))} \quad ,$$

T_1 being the set of pages containing t_1, T_2 being the set of pages containing t_2 (thus, $T_1 \cap T_2$ is the set of pages containing both terms), G is the total set of pages as reported by Google.

Showing less accuracy in comparison with a method describes in [23], Wikipedia link-based measure (WLM) has advantage of requiring less data and resources. Milne and Witted noticed that to obtain measures from ESA, huge amount of textual data must be preprocessed, while for WLM only the link structure of Wikipedia is required.

F. Approaches dealing with user experience

In [27] Ochoa and Duval introduced some metrics for relevance ranking for learning objects as special class of electronic documents. In learning space the user profile information plays bigger role in document ranking, because in learning every documents is being evaluated not only on the base of formal text similarity, but also on the base of subjective user's apprehension of the text: a document may be relevant to the query in regard of one user's experience but mostly irrelevant for other users. They introduced three classes of personalized relevance metrics – topical relevance metric, personal relevance metric and situational relevance metric, – and then adopted them to the leaning space tasks. The topical relevance metric and situational relevance metric are the most interesting within the scope of this study.

A basic topical relevance measure is aimed to estimate which objects (documents) are more related to what a given user wants. It is constructed as a sum of conceptual distances for selected objects:

$$BT(d,q)=\sum_{i=1}^{NQ} distance(q,q_i) \cdot selected(d,q_i)$$

$selected(d,q)=1, if\ d\ clicked\ in\ q$, d being the representation of document to be ranked, q being the user query, and q_i being the representation of the i-th previous query, NQ being the total number of queries. The distance between queries can be calculated either as the semantic distance between the query terms (by using WordNet or other ontology model), or the number of objects the both queries have returned in common.

The situational relevance measure is about estimating the relevance of the document to the specific tasks that caused the search. This measure is an adaptation of the $\text{sim}_{VSM}(d_1,d_2)$, presented earlier.

It seems that just presented measures may also be used for efficiency evaluation of some informational searching process. In other words, they can give us better understanding of how do assisting tools help users to achieve their searching goals.

Kerschberg et al. introduced a searching agent-based system that involve users in the process of searching by creating personalized taxonomy tree and by using user-specified weighting [28]. User intentions represented in a form of weighted taxonomy tree is then utilized to rank pages obtained from search engines classified into six matching categories. Each category (such as semantic matching, categorical matching, popularity ranks, and so on) may also be weighted in accordance to the user preference level.

It is clear that the usability requirements for those approaches, which rely on some supplementary weighting or term selection provided by users, demand special attention to the design of user interface components, so the complexity of the task is being shifted to the field of implementing flexible interfaces allowing users to participate in process of search more explicitly [29].

G. Using measures for summary generation

Creating summaries for the documents discovered on the Web by search engines is another example of an area, where relatedness measures are useful. General purpose search engines, as a rule, list the retrieved documents in form of short snippets. These snippets make evident the occurrence of query terms in the document, but don't provide indicative information about its contents [30].

The idea is to score document paragraphs or sentences to decide whether they should be included to the summary or not: sentences containing words that are the most relevant to key words are candidates to include into the summary.

To create summaries many traditional approaches described above may also be useful. But taking into account the fact that summary generation is not the first step in documents searching and selection in most cases (because the summaries are to be generated for documents, retrieved by a

engine, and therefore their appropriateness in respect to the user's needs is more or less guaranteed) the some simplified scoring techniques have been proposed and implemented. In [30], [31] there are examples of query-oriented approach to summarization based on the idea that different users may be interested in different aspects of the same document, so the summary that is good for one user may not be so descriptive for others. There is also wide range of works on generating query-independent summaries.

H. Combination measures

An evident way to improve relevance measures is to combine different approaches. One reason for this may be exposed by the citation from Budanitsky and Hirst work: "Computational applications typically require relatedness rather than just similarity" [4]. Another reason is a human way to deal with relatedness, since humans use rather a variety of approaches in combination, not only one approach.

Haralambous and Klyuev introduced the measure consisting of components related to the encyclopedic (based on Wikipedia concepts), ontological (WordNet) and collocational kind of knowledge [32].

For the encyclopedic component they used the ESA-measure introduced by Egozi, Markovitch and Gabrilovich in [24]. For the ontological component they studied different WordNet measures presented earlier in this paper and selected WordNet path measure showed better results when combined with ESA-measure.

To take into the consideration collocation nature of some word pairs in the word similarity test collections such as WS-353 [33] authors introduce collocation component.

The final expression for the combined EWC-measure is as follows:

$$\begin{aligned}\mathrm{sim}_{EWC}(w_1, w_2) = &\mathrm{sim}_{ESA}(w_1, w_2) \\ &\cdot (1 + \lambda \sigma_{(m,s)} \mathrm{sim}_{WNP}(w_1, w_2)) \\ &\cdot (1 + \lambda' \sigma_{(m',s')} C_{\xi}(w_1, w_2))\end{aligned}$$

where λ weights WNP with respect to ESA, m is sigmoid inflection point forming a soft boundary of WNP's lowest range, s is the steepness of the sigmoid, λ' m', and s' are similar to weight collocation component, and $C_{\xi}(w_1, w_2)$ is the mixed collocation index defined as follows:

$$C_{\xi}(w_1, w_2) = \frac{2 \cdot (freq(w_1 w_2) + \xi \cdot freq(w_2 w_1))}{freq(w_1) + freq(w_2)}$$

Values of λ, m, s, λ', m', s' and ξ are calculated numerically so to obtain highest value of Spearman ρ rank correlation coefficient between the relatedness ranking of pairs by human judges and that by the tested measure. Authors reported to achieve the value $\rho = 0.7874$ that, to the publication moment, is the highest result for WS-353 by a direct measure and only slightly inferior to the TSA-measure combining static semantic behavior of the words (achieved by ESA) with temporal dynamics introduced by Radinsky et al. in [34].

III. Invitation to the Discussion

In the above sections we reviewed mathematical approaches to measure document similarity. In this section we introduce some aspects of informational retrieval referring the human orientations of IR solutions since IR is a vast area of human centric computing. We mentioned earlier the term of research search aimed to discover the unclear domain, or to extend user experience. In this kind of search where the most important is not (only) how much the document is relevant, but how much is appropriate. Regardless that the relevance is a subjective judgment and may include aspects of subject, time, source authority, rank in the list of links, etc., if the search doesn't result the new knowledge, it isn't kind of successful search. This subjective nature of the relevance may lead to the following situation: a document may be considered relevant but it doesn't necessarily mean that the document is appropriate. The information may be adequate, but not so interesting for the user (for example, because this user already has got this information earlier).

It's ideal case, to rate relevance metrics according to their better correlation to human relevance estimation based on test collections of word pairs. There is some speculative component in estimating measures metrics by their concordance with human judgment on a given test suite. Evidently, metrics may fit the best the test suite, not the real searching problems.

In many measures, some "magic numbers" are being used (e. g. different constants, thresholds, correcting coefficients). Often they are introduced only as a result of some empirical research, in the best cases – as a solution of optimization task to obtain highest value of Spearman correlation.

For the day, it is still unclear how the Wikipedia (with its growing contents and refined concept taxonomy) implicitly affects the quality of measures.

There are open issues in regard to using cross-language information retrieval technologies, namely:

1. In searching, exploring documents in the languages different to the search query language may be useful in regards to the search comprehensiveness, but only in cases when the user may adequately conceive this document.
2. Cross-language retrieval may lead to the extending space of concepts necessary to relevance estimating, but it seems that there are only few works in this area (see [35] as an example). The discussion issue is whether the concepts relate to the language, or not.
3. Cross-language retrieval may be useful if there are too few document presented in the language of the query. For example, if we want to find information about some Japanese daimyo, in many cases the only way is to explore Japanese documents. But the problem of translating the query in another language and translating the response back to the user as well may seriously complicate the cross-language searching. Even the problem of name transcription may discourage searchers [35], [36].

IV. Conclusion

Ways measures work aren't always clear for humans. In their turn, relatedness measures are objects to evaluate as well. We compare measures to know which one works better

in the comparable conditions. Usually researchers use specially constructed collections of pairs of words or concepts which were firstly analyzed by human judges. This allows comparing approaches quite fairly, but doesn't take into the consideration certain aspects. One of them is relevance temporal dependency. In addition, measures may lead to judge relevance of "unexpected" documents with very complex implicit relations, which were rather not clear to humans. In [34] authors cited an example produced by the TSA approach with the pair "drink"/"car". For the reason that they used big corpus of New York Times as a source of temporal semantic information, and alcohol drinking is being mentioned often in connection with car accidents, "drink" and "car" words are being considered correlated, despite the fact the humans tend not to find them related.

In fact, this example doesn't illustrate the measure's weakness. The situation is rather normal: the discovered knowledge is new for the searcher, and the relatedness concluded on the base of relevance measures is in some way new for the user. That's why the user's feedback is of much importance to estimate measures in less formal way, but in more connection to real world searching problems. Hence, the task of developing some user feedback searching simulator that would enable to compare relevance measures dynamically seems promising.

Effectively, all the pairs in the WS-353 collection are well relevant, since they are included into the WS-353 set! Don't believe? Then ask Google.

References

[1] A.M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the web," ACM Transactions on Internet Technology, Vol.9, Issue 3 (July 2009), Article No.10.

[2] T. Saracevic, "Relevance reconsidered," In the Proceedings of the 2nd International Conference on Conceptions of Library and Information Science: Integration in Perspective (CoLIS2), Copenhagen, Denmark, 1996. P.Ingwersen and N.Pors Eds. The Royal School of Librarianship, pp.201-218.

[3] Ph. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural languages," Journal of Artificial Intelligence Research (JAIR), 11, 1999, pp.95-130.

[4] A. Budanitsky, and G. Hirst. "Evaluating WordNet-based measures of lexical semantic relatedness," Computational Linguistics, Vol. 32, No.1, March 2006, pp.13-47.

[5] A. Budanitsky. "Lexical semantic relatedness and its application in natural language processing: Technical Report CSRG-390," Computer Systems Research Group, University of Toronto, Aug.1999.

[6] B. Harrington, "A semantic network approach to measuring relatedness," In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010): Poster Volume, pp. 356-364. Beijing, August 2010.

[7] C. Leacock, and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," In "WordNet: An Electronic Lexical Database", C.Fellbaum, Ed., The MIT Press, Cambridge, Chapter 11, pp. 265-283.

[8] E. Altintas, E. Karsligil, and V. Coskun, "A new semantic similarity measure evaluated in word sense disambiguation," In the Proceedings of 15th NODALIBA conference, Joensuu, 2005, pp.8-11.

[9] Z. Wu, and M. Palmer, "Verb semantics and lexical selection,". In the Proceedings of the Symposium on Applications and the Internet, 2002 (SAINT'02), pp.230-237.

[10] G Hirst, and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," In "WordNet: An Electronic Lexical Database", C. Fellbaum, Ed., The MIT Press, Cambridge, Chapter 11, pp. 305-332.

[11] J.J. Jiang, and D Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," In the Proceedings of Internattional Conference on Computational Linguistics (ROCLING X), Taiwan, 1997, pp.19-33.

[12] T. Zesch, C. Mueller, and I. Gurevych, "Using wiktionary for computing semantic relatedness". In the Proceedings of the 23rd AAAI conference on artificial intelligence, AAAI 2008, Chicago, USA, 2008, pp.861-867.

[13] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic similarity methods in WordNet and their application to information retrieval on the Web," In the Proceedings of 7th annual ACM international workshop on Web information and data management(WIDM'05), Bremen, Germany, 2005, pp.10-16.

[14] G. Salton, "Automatic text processing: the transformation analysis and retrieval of information by computer," Addison-Wesley, 1989.

[15] A. Aizawa, "An Information-theretic perspective of TF-IDF measures," Information Processing and Management, vol.39, no.1, 2003, pp.45-65.

[16] V. Klyuev, and Ya. Haralambous, "Accurate query translation for Japanese-English cross-language information retrieval," In Proceedings of the International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS 2012), Rome, 2012, pp. 214-219.

[17] H. Kozima, and A. Ito, "Context-sensitive word distance by adaptive scaling of a semantic space," In R. Mitkov and N. Nicolov editors, "Recent advances in natural language processing: Selected papers from RANLP'95, vol.136 of "Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory", John Benjamins Publ. Comp., Amsterdam/Philadelphia, ch.2, pp.111-124.

[18] M. Baziz, M. Boughamen, N. Aussenac-Gilles, and C. Chrisment, "Semantic cores for representing documents in IR," In the Proceedings of the 2005 ACM Symposium of Applied Computing (SAC'05), Santa-Fe, New Mexico, USA, 2005, pp.1011-1017.

[19] M.E. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an Ice Cream Cone". SIGDOC, 1996.

[20] C. Mueller, and I. Gurevych , "A study on the semantic relatedness of query and document terms in information retrieval". In the Proceedings of the 2009 conference on empirical methods in natural langage processing, Singapore, 2009, pp. 1338-1347.

[21] J. Giles, "Internet encyclopedias go head to head," Nature, Dec 15, 2005.

[22] M. Strube, and S.P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," In Proceedings of the 21st national conference on Artificial intelligence, pp. 1419-1424. AAAI Press.

[23] E. Gabrilovich, and S. Markovitch, "Computing semantic relatedness using wikipedia based explicit semantic analysis". In the Proceedings of the 20th International joint conference for Artificial Intelligence, Hyderabad, India, 2007, pp. 1606-1611.

[24] O. Egozi, S. Markovitch, and E. Gabrilovich. "Concept-based information retrieval using explicit semantic analysis," ACM Transactions on Information Systems, Vol. 29, No. 2, Article 8, April 2011.

[25] D. Milne, and I. Witten, "An effective low-cost measure of semantic relatedness obtained from wikipedia links," In Wikipedia and AI workshop at the AAAI-08 conference (WikiAI08), Chicago, USA.

[26] R.L. Cilibrasi, and P.M.B. Vitanyi, "The Google similarity distance," IEEE Transactions on Knowledge and Data Engineering, Vol.19, Issue 3, 2007, pp.370-383.

[27] X. Ochoa, and E. Duval, "Relevance ranking metrics for learning objects," IEEE Transactions of Learning Technologies, Vol.1, No.1, Jan 2008, pp.34-48.

[28] L. Kerschberg, W. Kim, and A. Scime. "A personalizable agent for semantic taxonomy-based web search," In Lecture Notes on Artificial Intelligence. Springer, 3-31, 2003.

[29] E. Pyshkin, and A. Kuznetsov, "Approaches for web search user interfaces," FTRA Journal of Convergence, Vol.1, No.1, Dec. 15, 2010.

[30] V. Oleshchuk, and V. Klyuev, "Context-aware summary generation for web-pages," IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. Rende (Cosenza), Italy. September 21-23, 2009.

[31] E. Pyshkin, and V. Klyuev. "On document evaluation for better context-aware summary generation," In Proceedings of 2nd International Symposium on Aware Computing (ISAC2010), Nov. 1-4,

National Cheng-Kung University, Tainan, Taiwan, 2010. IEEE Catalog Number: CFP-1079K-CDR. ISBN: 978-1-4244-8312-9. CD Edition.

[32] Ya. Haralambous, and V. Klyuev, "A semantic relatedness measure based on combined encyclopedic, ontological and collocational knowledge," In Proceesdings of the 5th International Joint Conference on Natural Language Processing, pp. 1397-1402. Chiang Mai, Thailand November 8-13, 2011.

[33] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," ACM Transactions on Information Systems, 20(1): 116-131, 2002.

[34] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," WWW 2011, March 28 – April 1, 2111, Hyderabad, India.

[35] S. Hassan, and R. Mihalchea, "Cross-lingual semantic relatedness using encyclopedic knowledge." In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1192-1201. Singapore, August, 2009.

[36] V. Klyuev, and Ya. Haralambous, "A query expansion technique using the EWC semantic relatedness measure," Informatica 35 (2011) 401-406.