Knowledge and Information Systems manuscript No.
(will be inserted by the editor)

What's the Difference?

A Cognitive Dissimilarity Measure for Information Retrieval Result Sets

Carsten Keßler

Received: date / Accepted: date

Abstract Result rankings from context-aware information retrieval are inherently dynamic, as the same query can lead to significantly different outcomes in different contexts. For example, the search term *Digital Camera* will lead to different – albeit potentially overlapping – results in the contexts *Customer Reviews* and *Shops*, respectively. The comparison of such result rankings can provide useful insights into the effects of context changes on the information retrieval results. In particular, the impact of single aspects of the context in complex applications can be analyzed to identify the most (and least) influential context parameters. While a multitude of methods exists for assessing the relevance of a result ranking with respect to a given query, the question how *different* two result rankings are from a user's point of view has not been tackled so far. This paper introduces DIR, a cognitively plausible dissimilarity measure for information retrieval result sets that is based solely on the results and thus applicable independently of the retrieval method. Unlike statistical correlation measures, this dissimilarity measure reflects how human users quantify the changes in information retrieval result rankings. The DIR measure supports cognitive engineering tasks for information retrieval, such as work flow and interface design: Using the measure, developers can identify which aspects of context heavily influence the outcome of the retrieval task and should therefore be in the focus of the user's interaction with the system. The cognitive plausibility of DIR has been evaluated in two human participants tests, which demonstrate a strong correlation with user judgments.

Keywords Cognitive information retrieval · Human-computer interaction · Context awareness

C. Keßler
Institute for Geoinformatics
University of Münster
Germany
Tel.: +49–251–83–39764
Fax: +49–251–83–39763
E-mail: carsten.kessler@uni-muenster.de

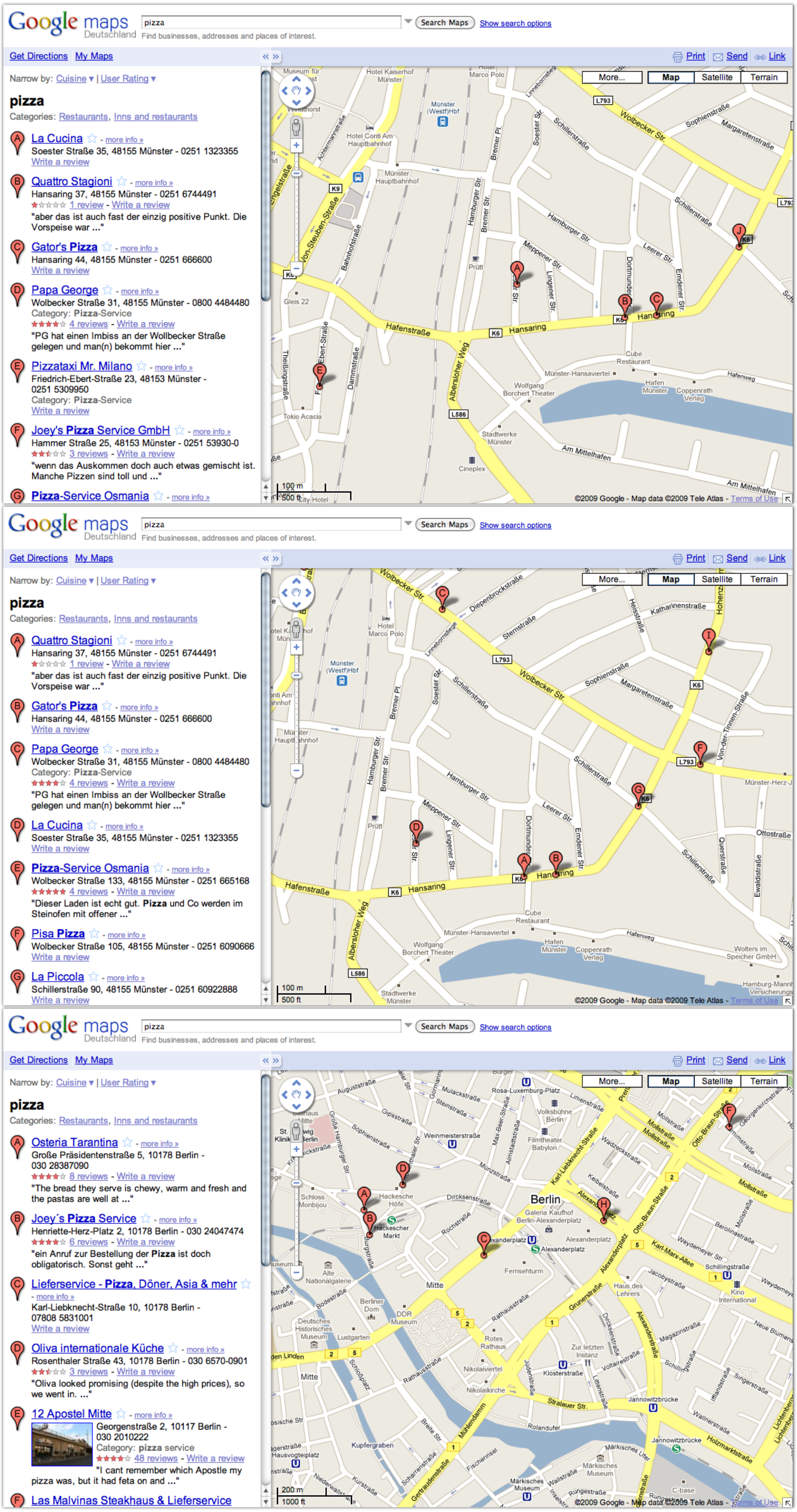
1 Introduction

User interaction with information retrieval (IR) results has been in the focus of research activities since the broad breakthrough of Web search engines. Complementary to existing approaches to assess the quality of search results for a given query [18], this research area concentrates on the presentation of results to the users and their interaction with them [3, chapter 10] and has been coined human-computer information retrieval (HCIR) [27]. A second trend within IR research is context-aware information retrieval (CAIR), focusing on the adaptation of IR results to personal interests and preferences, the user's current location and other context information relevant to the given task [6]. Changes of context can cause an adaptation of the result ranking for a given query, so that the same query does not necessarily always lead to the same result [13].

Consider the example of a search for *Pizza* on Google Maps shown in Figure 1. The top two screen shots show a map-based search in Münster, Germany, near central station, but with a slightly different map extent. The corresponding top results shown on the left have five matches in common, but in partly different order. The remaining matches are different in the two rankings. For the bottom screen shot, the spatial context has been moved to the area around Alexanderplatz in Berlin. Evidently, the results in the top two screen shots are very much alike, whereas the ranking in the bottom one is completely different from the other two. However, there is no existing measurement method that allows for the *quantification* of the difference between such result rankings from a user's perspective.

Comparisons of result rankings stemming from the same query posed in different contexts can provide useful insights into the effects of context changes on the IR results. While the context for the pizza example consists only of the selected map extent, applications such as a surf spot finder [21] require more detailed context models: In order to allow a user to retrieve a personalized selection of surf spots based on the current conditions, the context model must cover wind and water conditions, as well as the user's skills and preferences. In such a complex setting, an analytical tool that allows the developer to assess which context aspects can cause the biggest turbulences in the results when they change helps to reduce the complexity of the application by removing context aspects that have a negligible influence on the outcome. A measure for the *quantification* of the users' perception of the difference between two rankings would support the reduction of computational complexity of CAIR applications, as shown in Figure 2: Relevant context parameters can be identified if their change causes a significant modification of the results for a given query. For this purpose, the same query is posed repeatedly in different contexts, which differ in certain aspects. If the corresponding result rankings are modified beyond an application-dependent threshold, they have to be taken into account for the application. Otherwise, they can be neglected. This process can be automated for large numbers of queries and different constellations of context parameters to gain a detailed understanding of the influence of the different aspects of the context. In this way, such a measure would also support the development of simpler user interfaces for cognitive IR [45] by reducing the information shown to the user to relevant aspects of the task.

This paper introduces a cognitively plausible **D**issimilarity measure for **I**nformation Retrieval **R**esults (DIR) tailored to the specific requirements of this task. The measure analyzes the individual results in two rankings and compares them for overlap, focussing on the top results using a weighting mechanism to emphasize differences at



pizza

Narrow by: Cuisine ▾ | User Rating ▾

pizza

Categories: Restaurants, Inns and restaurants

A La Cucina ☆ - more info »
Soester Straße 35, 48155 Münster - 0251 1323355
Write a review

B Quattro Stagioni ☆ - more info »
Hansaring 37, 48155 Münster - 0251 6744491
★☆☆☆☆ 1 review - Write a review
"aber das ist auch fast der einzig positive Punkt. Die Vorspeise war ..."

C Gator's **Pizza** ☆ - more info »
Hansaring 44, 48155 Münster - 0251 666600
Write a review

D Papa George ☆ - more info »
Wolbecker Straße 31, 48155 Münster - 0800 4484480
Category: **Pizza**-Service
★★★★☆ 4 reviews - Write a review
"PG hat einen Imbiss an der Wollbecker Straße gelegen und man(n) bekommt hier ..."

E Pizzataxi Mr. Milano ☆ - more info »
Friedrich-Ebert-Straße 23, 48153 Münster - 0251 5309950
Category: **Pizza**-Service
Write a review

F Joey's **Pizza** Service GmbH ☆ - more info »
Hammer Straße 25, 48153 Münster - 0251 53930-0
★★★☆☆ 3 reviews - Write a review
"wenn das Auskommen doch auch etwas gemischt ist. Manche Pizzen sind toll und ..."

G **Pizza**-Service Osmania ☆ - more info »

pizza

Narrow by: Cuisine ▾ | User Rating ▾

pizza

Categories: Restaurants, Inns and restaurants

A Quattro Stagioni ☆ - more info »
Hansaring 37, 48155 Münster - 0251 6744491
★☆☆☆☆ 1 review - Write a review
"aber das ist auch fast der einzig positive Punkt. Die Vorspeise war ..."

B Gator's **Pizza** ☆ - more info »
Hansaring 44, 48155 Münster - 0251 666600
Write a review

C Papa George ☆ - more info »
Wolbecker Straße 31, 48155 Münster - 0800 4484480
Category: **Pizza**-Service
★★★★☆ 4 reviews - Write a review
"PG hat einen Imbiss an der Wollbecker Straße gelegen und man(n) bekommt hier ..."

D La Cucina ☆ - more info »
Soester Straße 35, 48155 Münster - 0251 1323355
Write a review

E **Pizza**-Service Osmania ☆ - more info »
Wolbecker Straße 133, 48155 Münster - 0251 665168
★★★★★ 4 reviews - Write a review
"Dieser Laden ist echt gut. **Pizza** und Co werden im Steinofen mit offener ..."

F Pisa **Pizza** ☆ - more info »
Wolbecker Straße 105, 48155 Münster - 0251 6090666
Write a review

G La Piccola ☆ - more info »
Schillerstraße 90, 48155 Münster - 0251 60922888
Write a review

pizza

Narrow by: Cuisine ▾ | User Rating ▾

pizza

Categories: Restaurants, Inns and restaurants

A Osteria Tarantina ☆ - more info »
Große Präsidentenstraße 5, 10178 Berlin - 030 28387090
★★★★☆ 8 reviews - Write a review
"The bread they serve is chewy, warm and fresh and the pastas are well at ..."

B Joey´s **Pizza** Service ☆ - more info »
Henriette-Herz-Platz 2, 10178 Berlin - 030 24047474
★★★★☆ 6 reviews - Write a review
"ein Anruf zur Bestellung der **Pizza** ist doch obligatorisch. Sonst geht ..."

C Lieferservice - **Pizza**, Döner, Asia & mehr ☆ - more info »
Karl-Liebknecht-Straße 10, 10178 Berlin - 07808 5831001
Write a review

D Oliva internationale Küche ☆ - more info »
Rosenthaler Straße 43, 10178 Berlin - 030 6570-0901
★★★☆☆ 3 reviews - Write a review
"Oliva looked promising (despite the high prices), so we went in. ..."

E 12 Apostel Mitte ☆ - more info »
Georgenstraße 2, 10117 Berlin - 030 2010222
Category: **pizza** service
★★★★☆ 48 reviews - Write a review
"I cant remember which Apostle my pizza was, but it had feta on and ..."

F Las Malvinas Steakhaus & Lieferservice

Figure 1 The spatial context for the query shown in the top two screen shots is very similar, leading to almost identical results for the search term *Pizza*. In the bottom screen shot, both the spatial context and the corresponding results are completely different from the top two.

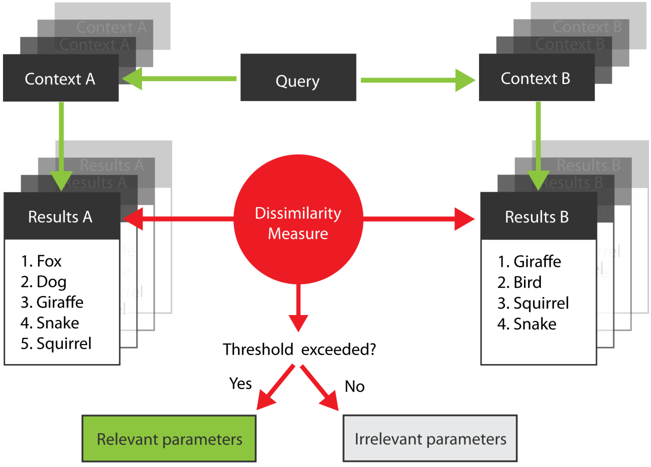


Figure 2 Identification of relevant context parameters.

the top ranks of the given result rankings. DIR returns values between 0 and 1, where 0 indicates that two rankings are completely identical, and 1 indicates that they are completely different, i.e., they share no common entries. Working solely on the results, the DIR measure can be applied independently of the actual IR method or search algorithm[1]. Rankings can be generated, for example, using the probability ranking principle [38] or novel approaches such as mean-variance analysis [52].

Statistical correlation measures [39] are for the most part not usable for the comparison of such result rankings. Beyond the fact that they have not been developed for inputs of different lengths and with potentially missing values on either side, which requires special handling [33], these measures also lack a mechanism to mimic the focus on the top results typically found in human behavior while searching for information [1]. Moreover, in order to reflect human judgments, it is essential that such a measure bears *cognitively plausible* [47] results, i.e., that the outcome it generates correlates strongly with human judgments.

To make sure that DIR measurements reflect the users' impression of changes in such IR result rankings, the measure's cognitive plausibility has been evaluated in two human participants tests. The first test was a controlled test with students from the University of California, Santa Barbara, where the participants were shown triplets of randomly generated result rankings. One ranking was used as a reference ranking. The participants' task was to rate which of the other two rankings differed more from the reference ranking. The second test was an open, Web-based test where the task was to rate the difference between two randomly generated rankings. The participants provided the ratings by positioning a slider between the two extremes *indistinguishable* and *no commonalities*. For both tests, the participants' judgments were statistically evaluated and show a strong correlation with the judgments calculated by the DIR measure.

The remainder of this paper is organized as follows: The next section introduces related work from the areas of context-aware and cognitive information retrieval as well

[1] In fact, the search algorithm may even be unknown, as it is the case for the example shown in Figure 1.

as statistical correlation measures. Section 3 introduces the DIR measure, followed by a description of the settings of the two human participants tests in Section 4. Section 5 evaluates the results of the tests and is followed by conclusions and an outlook on future work in Section 6.

2 Related Work

This section points to related work from the areas of cognitive and context-aware information retrieval as well as statistical correlation measures.

2.1 Context, Cognition, and Information Retrieval

Context is an ambiguous concept with numerous definitions and modeling approaches [4, 46]. Research on context is conducted in areas as different as pervasive computing, linguistics and sociology[2]. In computer science related fields, context is most frequently defined as follows:

Definition 1 *Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [7, p.5].*

While this definition generally also applies to IR, it does not tell whether a specific piece of information is really part of the context or not. Hence, we define context in IR as follows:

Definition 2 *An information retrieval task's context is any information whose change significantly modifies the task's outcome.*

This definition reduces context to relevant information and motivates the DIR measure, since it allows to determine whether a piece of information has a significant effect on the outcome of the information retrieval task or not.

Context has been successfully employed in a number of ways for personalization [22, 21] and to improve retrieval from the Web [9, 24], from email archives [53], from RDF graphs [31], from Wikipedia [49] as well as for ontology-based alert services [25]. Recent research has investigated how to reason about distributed contextual knowledge to enable ambient intelligence [5]. Different approaches based on vector spaces have been proposed to enable context-aware information retrieval [10, 36, 29]. Context, however, is only one part that contributes to cognitive IR [45], amongst interactive IR, feedback and explanations. With similar objectives, the term *human-computer information retrieval* (HCIR) was coined for a research field that combines information retrieval with human-computer interaction to open up new opportunities towards more intuitive, interactive, and goal-directed IR [41]. Marchionini [27] suggests "that integrating the human and system interaction is the main design challenge to realizing these opportunities and that what is required is design that recognizes a kind of syminforosis – people as organic information processors continuously engaged with information in the emerging cyberinfrastructure."

[2] For an overview, see the proceedings of the CONTEXT conference series [23].

Recently, semantic similarity measurement has gained attention as a tool for cognitive IR and HCIR. Originally a field of research in psychology [11], the artificial intelligence community started investigating the topic from a different perspective, developing specialized similarity measures [37, 2, 35, 56] and particularly focusing on how human similarity ratings can be mimicked by algorithmic approaches [37]. The dependence of human similarity judgments on the comparison task's context is a long-known phenomenon [28], which is also reflected in formal theories of similarity [43, 40, 15]. While these theories mostly define context by selecting a domain of application (also referred to as discourse context), recent research aims at a more comprehensive understanding of context by analyzing the context-aware similarity measurement process [19] and categorizing different kinds of context [14]. Kernel functions have been proposed to learn the appropriate similarity functions for nonlinear relationships among contextual information [55]. This paper builds on previous work aimed at a more general understanding of the effects of context changes on similarity measurements [20].

Previous research has investigated the origins of changes in Web search results. The PageRank algorithm [34] has been thoroughly analyzed in terms of the mathematical preconditions that need to be fulfilled in order to generate changes in rankings [30]. However, these observations are only valid for a specific algorithm. Moreover, they are not linked to the cognitive IR process in any way. By shifting the focus away from the retrieval method to the actual results, the DIR measure introduced in this paper can be used independently of the retrieval method.

2.2 Statistical Correlation Measures

Rank correlation is a well known problem with established statistical solutions. However, none of the standard correlation measures – such as Spearman's ρ [44], Kendall's τ [17] and Pearson's r [39] – are tailored to the specific requirements of comparing IR results. Generally speaking, this is due to the fact that those standard correlation coefficients have been developed to calculate correlation between sets of result pairs where every individual in one ranking has its corresponding counterpart in the other ranking. In information retrieval, however, cases where individuals (single result in this particular case) appear in one ranking, but not in the other, are very common. Thus, the two rankings to be compared often consist of different numbers of ranks. Finally, each rank may hold more than one match of equal relevance[3]. While this concept is hardly reflected in user interfaces for reasons of simplicity (see [15] for an example), IR algorithms mostly deal with such shared ranks internally. These ranks contain results that have been assigned the same relevance rating with respect to the current query. The DIR measure must be able to cope with all these kinds of input. Furthermore, the output of statistical correlation measures ranges from -1 for perfect negative correlation to 1 for perfect positive correlation. In the case of IR results, negative correlation is both hardly to expect and very difficult to explain. Hence, a mapping interval of $[0, 1]$ is more appropriate for the desired dissimilarity measure, where 0 applies for two identical result rankings and 1 applies for any pair of rankings that share no common entries.

From a cognitive perspective, statistical correlation coefficients present a purely mathematical view on the rankings at hand. While they can be employed to compare

[3] We use the term *relevance* here to refer to the calculated relevance of a hit based on the applied IR method, not for relevance feedback collected from users.

result rankings in principle (such as Kendall's τ in [8]), they do not reflect the human perspective on the results at hand. In particular, they lack the users' focus on the top results that is typical to information retrieval [1]. Rank correlation measures only compare the rankings for overlap, without any regard of *where* the differences between two rankings are – at the top or at the bottom. They therefore lack a mechanism to stress differences at the top ranks. This property, however, should be in the centre of a measure that supports cognitive IR. Comparable to the semantic similarity measures building on similarity research in psychology, the DIR measure needs to reflect the user's view on IR results. The cognitive aspect of this research is hence concerned with the question how to map from two given result rankings to a value between 0 and 1, so that the order of the resulting values corresponds to human judgments.

3 The DIR Measure

This section introduces the cognitively plausible dissimilarity measure for information retrieval results (DIR). The characteristics of the measure are discussed using examples from geographic information retrieval and the measure's formalization is introduced.

3.1 DIR Characteristics

During the development of tools for context-aware information retrieval, the question which aspects of the context to include is crucial from the users' as well as from a computational perspective. Users want the application to account for any *relevant* context changes, but processing power can be a bottle neck (especially on mobile devices) and the collection of context information can be costly when additional sensors are required. Furthermore, the computing of too much sensor input can slow down the application and hamper usability. Accordingly, it is important for application designers to be able to assess the impact of context changes to decide which aspects to include in their applications, and which to ignore. Finally, being able to measure the impact of context changes is also interesting from a theoretical perspective, since it enables distinguishing between noise and intended context [14]: intended contextual information must have an impact that goes beyond a threshold value δ, otherwise it is considered noise. When developers use DIR to pick the contextual aspects for a specific application, this selection should be evaluated using further analysis methods on the implemented system [48].

The DIR measure is purely based on result rankings. This context impact measure introduced in [20] in contrast calculates the changes in result relevance *values* for pairs of ontological concepts. Such a pair-based measure is of limited use, since it does not take a relevance value's interpretation context [14] into account, i.e., its relativeness regarding other values: A relevance value of 0.8 may result in a top rank, but depending on the other results, it can also be topped by a large number of other results with relevance values > 0.8, leading to a lower rank. Moreover, while a single relevance value (or even all values in a ranking) may change, the order within the ranking may still remain the same. Accordingly, we propose an approach that looks at result *rankings* instead of single relevance values. This ranking-driven approach also adheres to the cognitive aspect of IR, as the top of a ranking, presenting the best results for a specific query, is in the users' focus [1]. For DIR, this means that changes at the top of a ranking need to be emphasized by a higher weight, as opposed to changes that affect the lower ranks. The same applies to concepts that only appear in one of the two rankings at hand: they change the result rankings and disappear (and pop up, respectively) when going from one context to another. Both aspects are handled by a weighting mechanism introduced in Section 3.2.

DIR provides a tool that is applicable independently of the actual retrieval method, since the measure is based on the outcome of an information retrieval task in different contexts. Note that the measure is not intended to model – let alone explain – how human users compare two given rankings, but to reflect how human users quantify the differences between them. The focus of this research is on cognitive *plausibility*[4] rather than cognitive *adequacy*.

Statistical rank correlation measures are not applicable in information retrieval, as discussed in Section 2.2. They map to the interval $[-1, 1]$, where a value of 1 indicates identical rankings, -1 indicates perfectly inverse rankings, and 0 indicates no statistical correlation. DIR takes a different approach, mapping to the interval from 0 for equal rankings to 1 for the extreme case where no result appears in both rankings. The normalization to the interval $[0, 1]$ allows for a comparison of DIR measures independent of the length of the rankings and their properties. DIR does not produce any negative values since inverse rankings are not only hard to interpret, they are also very unlikely to occur in the comparison of information retrieval results. Moreover, statistical correlation measure lack the above-mentioned focus on the top results. Consider the following three rankings:

1.	apple	apple	orange
2.	mouse	mouse	mouse
3.	tree	tree	tree
4.	boat	boat	boat
5.	goat	ape	ape

Both the left and the right ranking differ in only one result from the ranking in the middle, which leads to the same rank correlation of 0.7 based on Spearman's ρ for both pairs of rankings. From a cognitive IR perspective, however, the left two rankings are much more alike than the right two rankings: while the former two only differ in the last result, the latter differ in the top result which is in the focus of the IR task.

Methods from information retrieval focus on measuring the quality of the results for a given query based on a collection of ground truth documents that have been classified as relevant in advance. The f-measure is a family of different combinations of recall and precision that is frequently applied for such quality assessments [50]. Other measures employed for the same purpose include the average normalized modified retrieval rank (ANMRR) employed in the MPEG-7 video standard [26]. ANMRR measures the retrieval quality over all queries for a collection of ground truth images, where retrieval is based on color and textures. While these measures also produce values between 0 and 1, they compare the contents of a ranking to a *non-ranked set* of ground truth items. They are hence not applicable to the comparison of two rankings.

An important aspect in human-computer information retrieval is the presentation of the results to the user. Flat, list-style visualizations of IR results that do not provide any information about the actual relevance of the shown results are the current

[4] What we refer to as cognitive plausibility in this paper has been defined as *relative* cognitive adequacy in [47].

standard, especially for Web search. Novel visualization techniques, such as tag clouds, give the user an impression of how good the results are, and how much better result 1 is than result 2, for example. Figure 3 shows two visualizations of a result ranking from geographic information retrieval: on the left, the results are presented without any indication of the relevance of the results. The tag cloud on the right indicates the relevance of the different results and the differences between them by font size: the better the result, the bigger the font used. Other visualization techniques also fall into these two categories, such as lists with relevance values or groupings of concepts into predefined categories [15]. An assumption that suggests itself is that the visualization of information about the relevance of results also influences the users' perception of changes in result rankings. In order to investigate whether this assumption holds, we have defined two different versions of the DIR measure: DIR$_{rank}$ is solely based on the position of the results in a ranking and is applied to list-style visualizations, whereas DIR$_{rel}$ takes the relevance values of the individual results into account. The mathematical definitions of the two variants of DIR will be introduced in the following subsection.

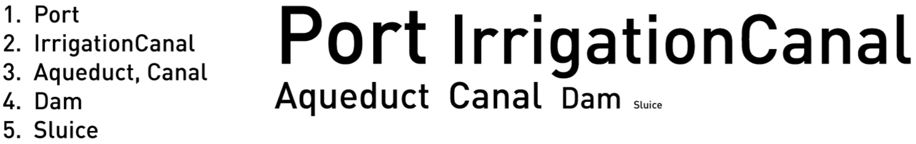


Figure 3 Visualization of the same result ranking as a list without indication of the relevance values (left) and as a tag cloud, where the relevance determines the font size (right).

3.2 Definition of DIR

DIR is defined based on the comparison of two result rankings. A result ranking R consists of an ordered set of ranks, where each rank consists of a relevance value $v \in [0, 1]$ and a non-empty set of results k. We assume that the ranks are in descending order with respect to the relevance values. Each rank is assigned an ascending rank number n, such that

$$\begin{aligned} R =& \langle \{1, v_1, (k_i, \dots, k_j)\}, \{2, v_2, (k_l, \dots, k_p)\}, \dots, \{n, v_n, (k_q, \dots, k_r)\}\rangle, \\ & \text{where } v_1 > v_2 > \dots > v_n. \end{aligned} \tag{1}$$

DIR is a symmetric function based on the *shift* that every concept k undergoes when a query is posed in two different contexts. This shift is calculated for every result that is part of either of the two rankings (or both). Based on the position in either ranking, individual weights are assigned to all shift values. The sum of all weighted shifts is then divided by the maximum possible dissimilarity *md* for normalization. By slight abuse of notation, we denote the appearance of a result k in a ranking M as $k \in M$. For a pair of result rankings M, N, we define DIR as

$$DIR(M, N) = \frac{\sum shift(k) * weight(k)}{md(M, N)}, k \in \{N \cup M\}. \tag{2}$$

In the following, we define the *shift*, *weight*, and *md* functions for the two variants DIR$_{rank}$ and DIR$_{rel}$.

The shift function. The basic formalization of DIR shown in eq. 2 can be filled with different functions depending on whether relevance values should be taken into account or not. The *shift* function for a purely rank-based DIR is the difference in rank number between the two rankings. A special case occurs when the result appears in only one of the two rankings: in this case, the shift is the distance between its current position and the position "behind" the last rank of the longer ranking. We hence treat the result as if it just drops out of the longer ranking[5]. Let $|M|$ be the number of ranks (not results) in M and $|N|$ the number of ranks in N. Assuming that M is the longer (or equal) ranking, i.e. $|M| \geq |N|$, we define the rank-based *shift* as

$$shift_{rank}(k) = \begin{cases} |rank_k(M) - rank_k(N)| & \text{if } k \in M \wedge k \in N, \\ |M| - rank_k(M) + 1 & \text{if } k \notin N, \\ |M| - rank_k(N) + 1 & \text{if } k \notin M. \end{cases} \quad (3)$$

The relevance-based *shift* function takes the same approach, though based on relevance values instead of ranks. It takes the relevance differences into account. The special case where a result appears in only one of the two rankings does not require special handling: results that do not appear in one ranking have a relevance value v of 0, so that the following equation covers all possible cases:

$$shift_{rel}(k) = |v_k(M) - v_k(N)| \quad (4)$$

The weighting function. The purpose of the weighting function is to stress shifts that affect the top of either ranking. DIR's weighting function, which is used both for DIR$_{rank}$ and DIR$_{rel}$, employs the rank number to determine the weight: if a result appears in both rankings, the rank number which is closer to the top of the ranking is used and subtracted from the number of ranks of the longer of the two rankings. This puts the weight in relation to the maximum possible weight for the given configuration (i.e., the weight for the results at rank 1 in the longer ranking). Accordingly, one could think of the weights as inverse rank numbers, i.e. in a ranking M, rank no. 1 is weighted $|M|$, rank no. 2 is weighted $|M| - 1$, and so forth. The last rank is weighted 1. If the result appears in only one ranking, the total number of ranks of the longer ranking is used as the weight. This stresses results that pop up or disappear during a context change, as explained in Section 3.1. For $|M| \geq |N|$, the weighting function is defined as

$$weight(k) = \begin{cases} 1 + max(|M| - rank_k(M), |M| - rank_k(N)) & \text{if } k \in M \wedge k \in N, \\ |M| & \text{if } k \notin M \vee k \notin N. \end{cases} \quad (5)$$

[5] Shifting a result "behind" its own ranking's last rank does not work here. For rankings with a big difference in length, this can produce situations where the shift of a result from the top of a short ranking to the end of a long ranking would exceed the *shift*$_{rank}$ value for dropping out, which is supposed to be the maximum possible shift.

Determining the maximum dissimilarity. According to Equation 2, the sum of all single impact values, $\sum \mathit{shift}(k) * \mathit{weight}(k)$, is then divided by the maximum possible dissimilarity value for normalization. The maximum dissimilarity md is reached when no result appears in both rankings. To calculate this value, we assume that this is the case for the given constellation of rankings with their specific relevance values and corresponding ranks. This maximum dissimilarity value depends on the shift function, therefore, the calculation of md is different for the rank-based and the relevance-based versions of DIR. For the rank-based version, this means that every concept is shifted out of the longer ranking and weighted according to the number of ranks in the longer ranking. Correspondingly, just like the shift function, the maximum possible impact depends on the lengths of the two rankings, and on the number of concepts at every rank. Let $k \in r_m$ be the concepts at rank m, $|k \in r_m|$ the number of concepts at rank m, and given $|M| \geq |N|$, the maximum impact for the rank-based DIR is

$$md_{rank}(M,N) = |M|*(\sum_{m=1}^{|M|}(|M|+1-m)*|k \in r_m|+\sum_{n=1}^{|N|}(|M|+1-n)*|k \in r_n|). \quad (6)$$

For the relevance-based DIR, this straight-forward calculation based on rank numbers is not possible. Instead, the relevance value at a given rank must be taken into account. Accordingly, md_{rel} additionally depends on the exact relevance values associated with the ranks. Let v_m be the relevance value at rank m, the maximum impact for the relevance-based DIR is

$$md_{rel}(M,N) = |M| * (\sum_{m=1}^{|M|} v_m * |k \in r_m| + \sum_{n=1}^{|N|} v_n * |k \in r_n|). \quad (7)$$

Both versions of md provide the maximum possible dissimilarity value for a given configuration of two relevance rankings. It is thus made sure that whenever two completely different result rankings are passed to DIR, it will return the value 1.

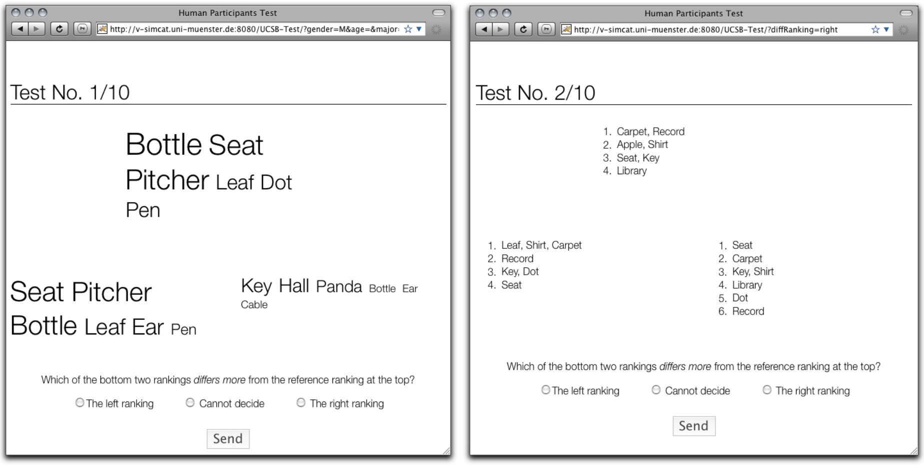
4 Human Participants Tests

The intention behind DIR is to represent how human users perceive changes in IR result rankings. This section describes two human participants tests that were carried out in order to evaluate the cognitive plausibility of the DIR measure.

4.1 Test One: Judgment by Order

The first human participants test (referred to as test one in the following) was a Web-based test designed around the task to compare two result sets to a third reference set (see Figure 4), where the participants had to rate which of the bottom two sets differs more from the reference set at the top. Three options were given to choose from: one could either select the left or right result set to be more different from the reference set, or alternatively choose "cannot decide". The first test hence focussed on the *order* of the different DIR values and its correlation with the order of the participant judgments [42, 16].

Though designed for completion on the Web, the test was not publicly accessible[6]. Login information was only provided for registered participants, who were shown an introductory page with an example to clarify the task. In some cases, the test was completed in a university lab; however, no additional instructions were provided. It was thus made sure that every participant receives the same information. Moreover, written instructions have been shown to be preferred by participants over spoken instructions [12].


Human Participants Test
http://v-simcat.uni-muenster.de:8080/UCSB-Test/?gender=M&age=&major:
Human Participants Test
http://v-simcat.uni-muenster.de:8080/UCSB-Test/?diffRanking=right

Test No. 1/10

Bottle Seat
Pitcher Leaf Dot
Pen

Seat Pitcher
Bottle Leaf Ear Pen

Key Hall Panda Bottle Ear
Cable

Which of the bottom two rankings *differs more* from the reference ranking at the top?

○ The left ranking ○ Cannot decide ○ The right ranking

Send

Test No. 2/10

1. Carpet, Record
2. Apple, Shirt
3. Seat, Key
4. Library

1. Leaf, Shirt, Carpet
2. Record
3. Key, Dot
4. Seat

1. Seat
2. Carpet
3. Key, Shirt
4. Library
5. Dot
6. Record

Which of the bottom two rankings *differs more* from the reference ranking at the top?

○ The left ranking ○ Cannot decide ○ The right ranking

Send

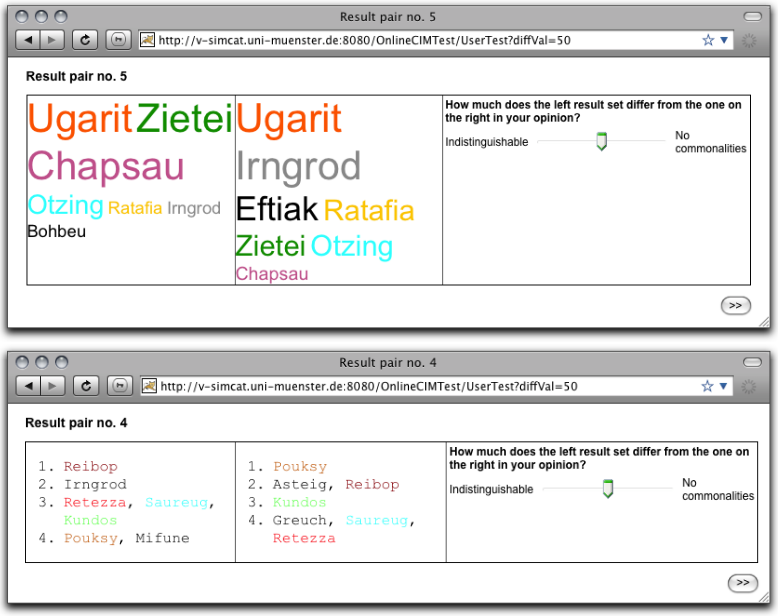
Figure 4 Screen shot of single comparison tasks in test one for the relevance-based (left) and rank-based (right) DIR variant.

Both visualization techniques were tested and every participant was shown all tasks using only one of them. The result rankings contained existing English words. For every participant, ten comparison tasks were randomly generated so that the ten difference values (*diff*) were equally distributed between 0 and 1: for every randomly generated triplet, $\text{DIR}_L = \text{DIR}(\text{LeftSet, ReferenceSet})$ and $\text{DIR}_R = \text{DIR}(\text{RightSet, ReferenceSet})$ were calculated, so that $\textit{diff} = |\text{DIR}_L - \text{DIR}_R|$. The *diff* values for the 10 tests shown to every participant were distributed randomly between 0 and 1. The hypothesis to test was that the bigger the *diff* value, the easier it should be for a participant to decide, reflected in higher numbers of judgments in line with DIR. Moreover, based on the logged time required for every task, it was assumed that tasks with a bigger *diff* value are solved in less time than those with smaller *diff* values, i.e., that there is a negative correlation between the time required per task and the *diff* value for the task.

[6] After the study, the first human participants test has been made publicly available for review at http://v-simcat.uni-muenster.de:8080/UCSB-Test/.

4.2 Test Two: Judgment by Numbers

The second test was also designed as a Web-based test[7], however, test two was an open test where everyone could participate. As in test one, participants were shown randomly generated result sets visualized either as tag clouds or lists (see Figure 5). The objective of this test was to investigate whether the DIR measure does not only correlate with participant judgments in terms of order, but also in terms of the actual values. The participants' task consisted in comparing pairs of result sets and answering the question "how much does the left result set differ from the one on the right in your opinion?". The questions were to be answered by moving a slider ranging from "indistinguishable" to "no commonalities". The position of the slider was internally mapped to values in the interval $[0, 1]$.



Result pair no. 5

http://v-simcat.uni-muenster.de:8080/OnlineCIMTest/UserTest?diffVal=50

Result pair no. 5

Ugarit Zietei Chapsau Otzing Ratafia Irngrod Bohbeu	Ugarit Irngrod Eftiak Ratafia Zietei Otzing Chapsau	How much does the left result set differ from the one on the right in your opinion? Indistinguishable — No commonalities

Result pair no. 4

http://v-simcat.uni-muenster.de:8080/OnlineCIMTest/UserTest?diffVal=50

Result pair no. 4

1. Reibop 2. Irngrod 3. Retezza, Saureug, Kundos 4. Pouksy, Mifune	1. Pouksy 2. Asteig, Reibop 3. Kundos 4. Greuch, Saureug, Retezza	How much does the left result set differ from the one on the right in your opinion? Indistinguishable — No commonalities

Figure 5 Screen shots of single comparison tasks in test one for the relevance-based (top) and the rank-based (bottom) DIR variants.

The test was available in English and German to increase the number of potential participants. It was explained with an example and some background information on an introductory Web page. The participants were specifically advised that there are no right or wrong answers in the test to make sure that they adjust the slider according to their personal opinion. In addition to the slider position, the time every participant spent on every pair of result rankings until he or she committed the judgment was logged. Each ranking of test data was stored along with the computer's IP address to make sure that every participant takes the test only once.

[7] The test is available online at http://v-simcat.uni-muenster.de:8080/OnlineCIMTest/UserTest?language=en.

As in test one, every participant was presented with all comparison tasks using the same visualization technique – either list-based or as a tag cloud – to avoid confusion. The result rankings were limited to seven entries on either side to prevent cognitive overload. As opposed to test one, the results in the rankings consisted of made-up words, so that the participants could not perform any undesirable reasoning on the result rankings, such as "*result set A contains 'tree' and result set B contains 'apple', therefore...*". Moreover, these made up words guaranteed that the German and the English speaking participants could be shown the same test data. The entries were also color-coded to facilitate recognition of identical entries in the two result sets. Entries appearing in both rankings were displayed in the same color, whereas entries shown only either left or right were displayed in black.

The seven pairs of result sets were arranged according to predefined ranges using DIR. Every participant was shown two extreme cases, one with completely identical rankings (DIR = 0) and one with completely different rankings (no common entries, DIR = 1). These two cases were included to make sure that every participant had read and understood the instructions. The remaining five were generated so that the corresponding DIR values were within five equal intervals between 0 and 1 (i.e., $[0.0, 0.2[, [0.2, 0.4[\ldots [0.8, 1.0]$). The seven tests were randomly generated for every new participant and presented in random order.

5 Evaluation and Comparison

This section analyzes and compares the results of the two human participants tests introduced in Section 4. The distinction between the two different visualization types is discussed.

5.1 Test One: Judgment by Order

The participants in test one consisted of a group of 52 undergraduate students at the University of California, Santa Barbara, who received credit points for participation. The participants were between age 17 and 27, with a mean age of 18.5. The group consisted of 37 female and 14 male[8] participants majoring in different fields, most of them still undecided about their major.

The evaluation of test one was carried out on the complete collection of all 52 participant datasets. Figure 6 gives an overview of the collected data and shows that the cases where the participant judgements coincide with the calculated *diff* value based on DIR form the large majority. Nonetheless, some participants largely disagree with the calculated values, such as participant 02 in the list visualization, or participant 05 in the tagcloud visualization. Out of the 520 comparison tasks completed by the participants, 341 ($\sim$ 66%) were judged in compliance with DIR, 111 ($\sim$ 21%) were judged inconsistent with DIR, and in 68 cases ($\sim$ 13%), the participants could not decide.

Figure 7 shows the distribution of 'correctly' (i.e., in line with DIR), 'falsely' (contradicting DIR) and undecided tasks per user. The right histogram shows that only 5 participants ($\sim$ 10%) judged more than 4 tasks contradicting DIR; for comparison, 22

[8] One participant did not provide any personal information.

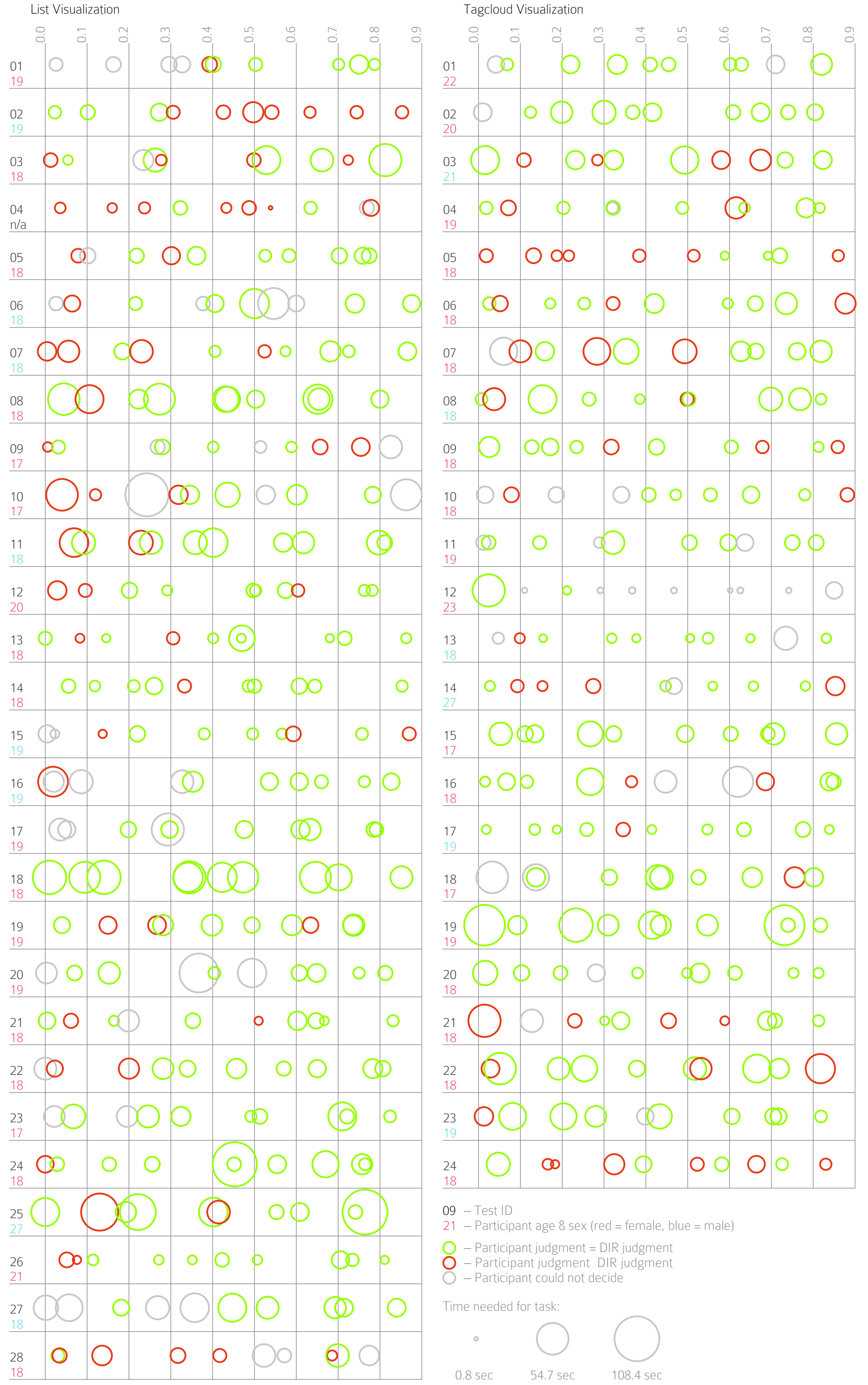


Figure 6 Results for the 52 participants in test one. Every row in the two columns represents one participant, with one circle per comparison task. The circle's color indicates whether the user judgment matches DIR, the circle area reflects the time the participant needed to answer. The circles' positions in the row reflect the respective *diff* values.

participants ($\sim$ 42%) judged less than 2 tasks contradicting DIR. The left histogram shows that 37 participants ($\sim$ 71%) judged more than half of the tasks according to DIR. These numbers indicate a strong agreement of the participants with the DIR measure, which is supported by the results of a test for correlation. For this test, all 520 tasks completed by the participants were sorted with ascending *diff* values and then split into 20 groups of 26 tasks each. For each group, the mean *diff* value and the number of participant judgments that were in line with DIR were computed. Since the resulting ranks are not equidistant, the correlation between mean *diff* value and number of 'correctly' solved tasks was tested using Kendall's τ. The test yields a positive correlation of $\tau = 0.602$. The result is significant at the 1% level (p= 0.000029). Hence, the hypothesis that higher differences between the DIR values lead to a higher recognition rate among the participants cannot be falsified (i.e., we have to reject the null hypothesis of no correlation).

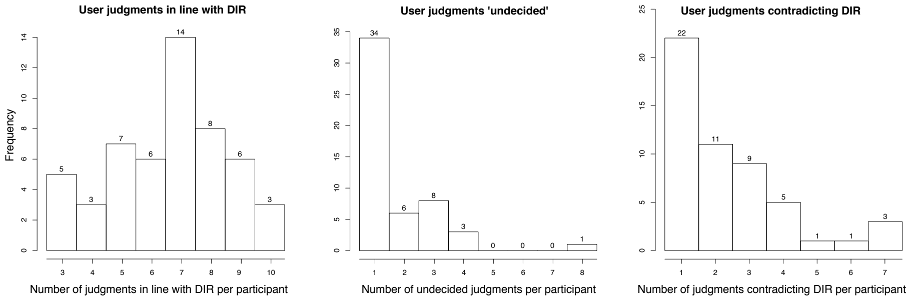


Figure 7 Histograms of the number of tasks solved in line with DIR (left), undecided tasks (middle) and tasks solved contradicting DIR (right) per participant. In the left histogram, for example, the rightmost bar shows that 3 participants had solved all 10 tasks in line with DIR.

The second hypothesis to test was that a higher *diff* value in a specific task also allows the participants to solve it faster. Looking for a negative linear correlation between the respective *diff* values and logged completion times, this hypothesis was tested using Pearson's product-moment correlation on the 520 single tasks. The test yields a slight negative correlation of $\rho = -0.11$, which is significant at the 5% level (p= 0.0123). Accordingly, the *diff* values have a comparably small (yet statistically significant) effect on the time required to solve a task.

The development of the two different versions of DIR followed the assumption that user judgements depend on whether the result visualization includes hints on inter-rank distances or not. If this is the case, the correct choice between DIR_{rel} and DIR_{rank} with respect to the presentation of the results should increase the correlation between user judgements and the DIR calculations. To test this hypothesis, the two correlation tests were repeated with a dataset where the *diff* values for all tasks carried out by the participants were calculated using DIR_{rank}. The DIR values were thus calculated only based on the order of the results, independent of whether the participants were shown tag clouds containing information on inter-rank distance. In this setting, Kendall's τ yields a positive correlation of $\tau = 0.613$, significant at the 1% level (p= 0.00020). The correlation is thus even slightly stronger when DIR is calculated solely based on ranks, which suggests that DIR_{rank} is a plausible measure for perceived changes in result sets, independent of result visualization. Accordingly, inter-rank distance seems to play a secondary role for users, who seem to focus on the order of the results. The correlation between *diff* values and completion times remains largely unchanged by the limitation to DIR_{rank}.

5.2 Test Two: Judgment by Numbers

The 81 participants who took the test were anonymous volunteers recruited by distributing an announcement on several mailing lists. As no compensation was granted for participation, the test was intentionally kept very short to avoid participants from getting bored and breaking off in the middle of the test. No personal data such as age, sex, or occupation were collected, as the reliability of such data in an open, Web-based human participants test is per se very limited. It is very likely that the majority are adults with a college education background, given that the call was distributed on several geography, geoinformatics and computer science mailing lists.

Out of the 81 participants who took test two, 42 were shown the rank-based test and 39 were shown the relevance-based test. Only tests where the participants had correctly understood the instructions were taken into account for the statistical analysis: if one of the two extreme cases contained in each test ($\mathrm{DIR} = 0$ or $\mathrm{DIR} = 1$) were judged more than 0.25 “off” (user judgment > 0.25 and < 0.75, respectively), these tests were not used any further. After this filtering, 31 rank-based and 25 relevance-based tests were used for evaluation. For this collection of 56 tests, the participants spent a mean time of 4 min, 16 sec on the completion (including instructions) and a mean time of 2 min, 14 sec on reading the instructions. It it thus safe to assume that the tests used for evaluation were completed thoroughly and that the according participants had read the instructions and understood the task.

Figure 8 shows bubble plots of DIR values against the values selected by the participants using the slider. The circle area reflects the time spent on the respective task. The distribution of the bubbles, each representing the result of one task completed by a participant, already gives an impression of the high variance of the user judgments compared to the calculated DIR values (see also Figure 9). Nonetheless, the regression line in Figure 8 shows that the calculated DIR values generally follow the same trend as the participant values.

The correlation between the user judgments and computed DIR values for test two can be calculated directly on the values entered by the users; it is not necessary to group the tasks, as in the evaluation of test one. When both visualization types are analyzed with the respective DIR variant, Pearson’s product-moment correlation yields a positive correlation of 0.805, significant at the 1% level. In this constellation, the logged completion time correlates slightly with the respective DIR values at 0.106, significant at the 5% level. Interestingly, the users thus needed slightly longer, the more different (according to DIR) the two shown rankings were.

For test two, we also analyzed the effect of the choice of DIR method on the correlation. If all results are recalculated using only DIR_{rank}, the correlation between user judgments and DIR values decreases slightly to 0.803 (significant at the 1% level). The correlation between DIR values and completion times does not change notably either, dropping to 0.102.

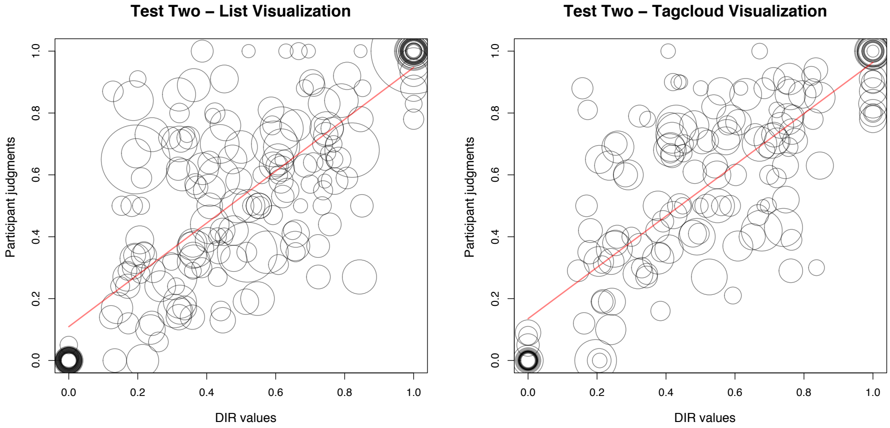


Figure 8 Bubble plots with regression lines of the calculated DIR values against the participant judgments for list visualization (left, with DIR$_{rank}$) and tag cloud visualization (right, with DIR$_{rel}$) in test two. The bubble size indicates the logged completion time for the respective task.

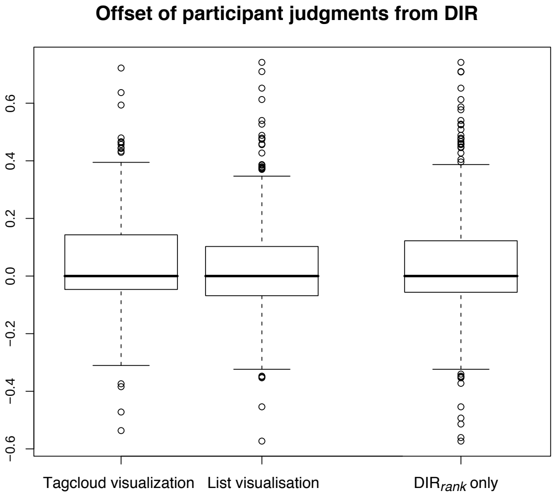


Figure 9 Box plots of the two variants of test two, showing the deviation of the participant judgments from the computed DIR values. The two box plots on the left show both visualization types with their corresponding DIR variants, the right box plot shows all 392 tasks solved by the participants compared to DIR$_{rank}$. The boxes are slightly shifted to the positive side, which shows that the participants tend to rate the dissimilarity slightly higher than DIR.

5.3 Discussion

The most striking outcome of the evaluation is the fact that the purely rank-based variant of DIR correlates equally strong with the participant judgments as the relevance-based DIR in both tests. Apparently, users do not take the relevance values of the results into account when they compare two sets. While the information that, for example, result 1 is twice as good as result 2, is certainly valuable from an objective point of view, users do not seem to distinguish this from a case where result 1 is only slightly better than result 2. The outcome of the two human participants tests thus confirms the approach taken by all popular Web search engines in presenting their results to the users as flat lists. At the same time, it renders DIR$_{rel}$ superflous, at least for the cases studied in this research. This may change when applying the measure to real IR use cases, where users typically only skim through the results and do not take such a close look as the participants in these two studies.

The evaluation of both human participants tests revealed a strong correlation between the calculated DIR values and the respective participant judgments. Completion time, however, does not seem to be influenced by how different (according to DIR) the presented result sets are. A slight negative correlation could be observed in test one, showing that the more different two pairs of result sets, the faster the tasks were solved. The evaluation of test two did not support this finding, though. Since in both cases, the correlation was very weak, we have to assume that there is no connection between DIR and processing times – at least at the level of number of results presented in this study. A stronger correlation may turn out with higher numbers of results presented to the participants. Moreover, this outcome may have been influenced by the setting of the study, where all participants looked thoroughly at the tasks before submitting their judgments; to further investigate this correlation, observing users in solving actual information retrieval tasks seems to be a promising approach.

6 Conclusions and Future Work

For the development of context-adaptive applications for information retrieval, tools for the analysis of the impacts of context changes are indispensable. In this paper, we have introduced DIR, a cognitively plausible dissimilarity measure for information retrieval result sets. Two different versions of DIR were introduced, depending on the visualization of the results. Both variants – one for purely rank-based visualizations (DIR$_{rank}$) and one for visualizations containing information on result relevance (DIR$_{rel}$) – were tested in two human participants tests. The evaluation of the tests showed that there is a strong correlation between the results calculated by the DIR measure and the participant judgments, proving the cognitive plausibility of the measure. In both cases, DIR$_{rel}$ was outperformed by DIR$_{rank}$ in terms of correlation. Besides the DIR measure, the major contribution of this paper is hence to show that users seem to focus on the order of the results and largely ignore the relevance information when comparing result sets. Moreover, the evaluation did not show any evidence of a correlation between the difference of two result sets and the time required to compare them. In the settings of the studies, i.e., with a comparably low number of results, it did thus not matter for the completion time whether the results at hand were very much alike or very different. Previous research indicates, however, that the small number of results presented to the participants were still within a range that can easily be processed [32]. It thus remains to be shown whether this finding also holds when the participants are shown larger numbers of results.

Future research should hence focus on testing these findings in studies where users are solving actual information retrieval tasks. In the human participants tests discussed in this paper, the result sets presented to the users were reduced to small numbers of results. Moreover, the participants' task was to compare the different result sets –

in a more realistic scenario, users would have to solve an actual IR task with larger numbers of results. Conclusions on their comparison judgments would then have to be drawn from their interaction with the system. Potential testing scenarios are especially attractive in controlled environments with directs access to the users. Such conditions are given for information systems in museums or libraries, for example, where visitors search the respective databases and catalogues. This setup would also allow for interviews with the users, either during or after use of the information system, to find out about their strategies for comparing result rankings. Concerning mobile applications, smartphone implementations of tools such as a surf spot finder [21] or a climbing route planner [54] are attractive for user studies because of their sensitivity to context changes. Since the evaluation of mobile applications can be challenging, participants could be ask to "think aloud" during their interaction with the system. The built-in microphone could be used to record the users' utterances. Besides these studies which analyse the users' interaction with IR systems during actual use, one could also imagine eye tracking studies to derive from the users' eye movements how they compare rankings. In any case, such studies should also broaden the range of participants in terms of background, education and experience in using Web search engines. For both tests presented in this paper, the participants can be assumed to be experienced Web users who are familiar with IR systems, at least in the form of Web search engines. Test one was even completed by a group of participants consisting only of *digital natives* who have grown up with the Web. It would thus be interesting whether their results can be confirmed for less experienced users.

The DIR measure works solely on the result sets, independent of the applied IR method. As such, it is also applicable to analyze the effects of changes to knowledge bases or search engine indexes. In peer-to-peer based search engines [51], DIR can investigate how nodes joining or leaving the network affect search result. For context-aware information retrieval, DIR allows us to *quantify* the changes in the result sets during context changes. The *qualitative* aspect of these changes in the results can simply be identified by looking at how the single results moved up or down in the ranking, and by checking which of them disappeared or turned up. The final missing piece in fully analyzing the influence of context changes on result sets is thus the link between the qualitative changes in context and the qualitative changes in the results. The goal of such an analysis is to be able to predict the changes in the results that follow from changes in specific aspects of the context. Whether this is possible independently of the applied IR method remains an open research question.

Acknowledgements This work is funded by the *SimCat* project granted by the German Research Foundation (DFG Ra1062/2-1 and Ja1709/2-2), see http://sim-dl.sf.net. Special thanks go to Dan Montello and Marius Austerschulte for their help with the human participants tests, and to Martin Raubal and Werner Kuhn for valuable comments and discussions.

References

1. E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10. ACM New York, NY, USA, 2006.

2. R. Albertoni and M. De Martino. Asymmetric and context-dependent semantic similarity among ontology instances. *Journal on Data Semantics X*, Lecture Notes in Computer Science 4900:1–30, 2008.
3. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.
4. M. Bazire and P. Brézillon. Understanding Context Before Using It. In A. K. Dey, B. Kokinov, D. Leake, and R. Turner, editors, *Modeling and Using Context – 5^{th} International and Interdisciplinary Conference (CONTEXT 2005), Paris, France*, volume 3554 of *Lecture Notes in Computer Science*, pages 29–40. Springer-Verlag Berlin Heidelberg, 2005.
5. A. Bikakis, G. Antoniou, and P. Hasapis. Strategies for contextual reasoning with conflicts in ambient intelligence. *Knowledge and Information Systems*, April 2010.
6. P. J. Brown and G. J. F. Jones. Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. *Personal and Ubiquitous Computing*, 5:253–263, 2001.
7. A. Dey. Understanding and Using Context. *Personal Ubiquitous Computing*, 5(1):4–7, February 2001.
8. M. Efron. Using multiple query aspects to build test collections without human relevance judgments. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 276–287. Springer-Verlag, 2009.
9. L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: the concept revisited. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 406–414, New York, NY, USA, 2001. ACM Press.
10. P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* MIT Press, 2000.
11. R. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning.* Cambridge University Press, 2004.
12. S. Harrison. A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 82–89, New York, NY, USA, 1995. ACM Press/Addison- Wesley Publishing Co.
13. J. Hu and P. Chan. Personalized web search by using learned user profiles in re-ranking. In *Workshop on Knowledge Discovery on the Web, KDD conference*, pages 84–97, 2008.
14. K. Janowicz. Kinds of Contexts and their Impact on Semantic Similarity Measurement. In *5^{th} IEEE Workshop on Context Modeling and Reasoning (CoMoRea) at the 6^{th} IEEE International Conference on Pervasive Computing and Communication (PerCom'08)*, 2008.
15. K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Bäumer. Algorithm, Implementation and Application of the SIM-DL Similarity Server. In F. Fonseca and M. Rodríguez, editors, *Second International Conference on GeoSpatial Semantics, GeoS 2007*, volume 4853 of *Lecture Notes in Computer Science*, pages 128–145. Springer-Verlag Berlin Heidelberg, 2007.
16. K. Janowicz, C. Keßler, I. Panov, M. Wilkes, M. Espeter, and M. Schwarz. A Study on the Cognitive Plausibility of SIM-DL Similarity Rankings for Geographic Feature Types. In L. Bernard, A. Friis-Christensen, and H. Pundt, editors, *The European Information Society—Taking Geoinformation Science One Step Further*

(AGILE 2008 Proceedings), Lecture Notes in Geoinformation and Cartography, pages 115–134. Springer-Verlag Berlin Heidelberg, 2008.

17. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(12), 1938.
18. A. Kent, M. M. Berry, J. Fred U. Luehrs, and J. W. Perry. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93–101, 1955.
19. C. Keßler. Similarity Measurement in Context. In B. Kokinov, D. Richardson, T. Roth-Berghofer, and L. Vieu, editors, *6^{th} International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark*, volume 4635 of *Lecture Notes in Artificial Intelligence*, pages 277–290. Springer-Verlag Berlin Heidelberg, 2007.
20. C. Keßler, M. Raubal, and K. Janowicz. The Effect of Context on Semantic Similarity Measurement. In R. Meersman, Z. Tari, and P. Herrero, editors, *On The Move – OTM 2007 Workshops, Part II*, volume 4806 of *Lecture Notes in Computer Science*, pages 1274–1284. Springer-Verlag Berlin Heidelberg, 2007.
21. C. Keßler, M. Raubal, and C. Wosniok. Semantic rules for context-aware geographical information retrieval. In P. Barnaghi, K. Moessner, M. Presser, and S. Meissner, editors, *Smart Sensing and Context, 4th European Conference, EuroSSC 2009, Guildford, UK, September 2009*, volume 5741 of *Lecture Notes in Computer Science*, pages 77–92. Springer-Verlag Berlin Heidelberg, 2009.
22. H. R. Kim and P. K. Chan. Learning implicit user interest hierarchy for context in personalization. *Applied Intelligence*, 28(2):153–166, 2008.
23. B. Kokinov, D. Richardson, T. Roth-Berghofer, and L. Vieu, editors. *Modeling and Using Context 6^{th} International and Interdisciplinary Conference CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007, Proceedings*, volume 4635 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag Berlin Heidelberg, 2007.
24. R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar. Searching with context. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 477–486, New York, NY, USA, 2006. ACM Press.
25. A. Leonidis, G. Baryannis, X. Fafoutis, M. Korozi, N. Gazoni, M. Dimitriou, M. Koutsogiannaki, A. Boutsika, M. Papadakis, H. Papagiannakis, G. Tesseris, E. Voskakis, A. Bikakis, and G. Antoniou. Alertme: A semantics-based context-aware notification system. In *33rd Annual IEEE International Computer Software and Applications Conference*, pages 200–205. IEEE, 2009.
26. B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703–715, 2001.
27. G. Marchionini. Toward human-computer information retrieval. June/July 2006 Bulletin of the American Society for Information Science, available online at http://www.asis.org/Bulletin/Jun-06/marchionini.html, 2006.
28. D. Medin, R. Goldstone, and D. Gentner. Respects for Similarity. *Psychological Review*, 100(2):254–278, 1993.
29. M. Melucci. A basis for information retrieval in context. *ACM Transactions on Information Systems*, 26(3):1–41, 2008.
30. M. Melucci and L. Pretto. PageRank: When order changes. Lecture Notes in Computer Science 4425, pages 581–588. Springer-Verlag Berlin Heidelberg, 2007.
31. B. A. Meza, C. Halaschek, B. I. Arpinar, and A. Sheth. Context-aware semantic association ranking. In *Semantic Web and Databases Workshop Proceedings*, pages 33–50, Berlin, Germany, September 2003.

32. G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.
33. K. A. Nedas and M. J. Egenhofer. Integral vs. separable attributes in spatial similarity assessments. In *Proceedings of the international conference on Spatial Cognition VI*, pages 295–310. Springer-Verlag Berlin Heidelberg, 2008.
34. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
35. D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(3):361–394, 2009.
36. M. Raubal. Formalizing Conceptual Spaces. In L. Vieu and V. A., editors, *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, Frontiers in Artificial Intelligence and Applications, pages 153–164. IOS Press, Amsterdam, NL, 2004.
37. E. Rissland. AI and Similarity. *IEEE Intelligent Systems*, 21(3):39–49, 2006.
38. S. E. Robertson. The probability ranking principle in ir. pages 281–286, 1997.
39. J. L. Rodgers and W. A. Nicewanderer. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.
40. A. Rodríguez and M. J. Egenhofer. Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, 18(3):229–256, 2004.
41. D. E. Rose and D. Levinson. Understanding user goals in web search. In S. Feldman and M. Uretsky, editors, *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM Press, 2004.
42. S. Rosset, C. Perlich, and B. Zadrozny. Ranking-based evaluation of regression models. *Knowledge and Information Systems*, 12(3):331–353, 2007.
43. A. Schwering. Approaches to Semantic Similarity Measurement for Geo-Spatial Data—A Survey. *Transactions in GIS*, 12(1):5–12, 2008.
44. C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15:72–101, 1904.
45. A. Spink and C. Cole, editors. *New Directions in Cognitive Information Retrieval.* Springer Netherlands, 2005.
46. T. Strang and C. Linnhoff-Popien. A context modeling survey. In *First International Workshop on Advanced Context Modelling, Reasoning And Management at UbiComp 2004, Nottingham, England, September 7, 2004*, 2004.
47. G. Strube. The role of cognitive science in knowledge engineering. In *Proceedings of the First Joint Workshop on Contemporary Knowledge Engineering and Cognition*, volume 622 of *Lecture Notes In Computer Science*, pages 161–174. Springer, 1992.
48. L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Knowledge and Information Systems*, July 2009.
49. A. Ukkonen, C. Castillo, D. Donato, and A. Gionis. Searching the wikipedia with contextual information. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 1351–1352, New York, NY, USA, 2008. ACM.
50. C. J. van Rijsbergen. *Information Retrieval.* Butterworth, 2 edition, 1979.
51. D. Wang, Q. Tse, and Y. Zhou. A decentralized search engine for dynamic web communities. *Knowledge and Information Systems*, December 2009.

52. J. Wang. Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 4–16. Springer-Verlag, 2009.
53. W. Weerkamp, K. Balog, and M. de Rijke. Using contextual information to improve search in email archives. In *Advances in Information Retrieval. 31st European Conference on Information Retrieval Conference (ECIR 2009)*, pages 400–411. 2009.
54. M. Wilkes. A graph-based alignment approach to context-sensitive similarity between climbing routes. Diploma thesis, Institute for Geoinformatics, University of Münster, Germany, 2008.
55. G. Wu, E. Y. Chang, and N. Panda. Formulating context-dependent similarity functions. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 725–734, New York, NY, USA, 2005. ACM.
56. J. Yang, W. Cheung, and X. Chen. Learning element similarity matrix for semi-structured document analysis. *Knowledge and Information Systems*, 19(1):53–78, April 2009.