Graphical Representation and Similarity Measurement of Relevance Judgments on the Web

Panos Balatsoukas
Department of Computer and Information Sciences
University of Strathclyde
26 Richmond Street, Glasgow, G1 1XH, UK
+44 141 548 3092
panos@cis.strath.ac.uk

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
26 Richmond Street, Glasgow, G1 1XH, UK
+44 141 548 3098
ian.ruthven@cis.strath.ac.uk

ABSTRACT
The purpose of this paper is to present a method for the graphical representation and similarity measurement of relevance judgments on the web. In order to address this objective a Latent Semantic Indexing technique was used. The findings suggest that the proposed method could help researchers in information seeking and retrieval to make methodological decisions about their data, such as the selection of specific subsets of relevance judgments for further examination, the recording of dissimilarities between judgments, or, the identification of possible cognitive shifts and abnormalities in relevance judgment behavior during web searching.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval-Selection Process.

General Terms
Human Factors, Measurement.

Keywords
Relevance judgment behavior, User studies.

1. INTRODUCTION
To date, the cognitive approach to the study of relevance judgment behavior on the web has been focused on the identification of relevance criteria (such as Topicality, Scope, Tangibility and Quality) and the use of descriptive statistics in order to explain their level of uptake during the web search process (such as the frequency of use of these criteria) [1]. Few researchers have gone a step further and tried to provide predictive models of relevance criteria use [2], or, map their association with interface components on the web [3]. Although these studies have advanced our understanding of relevance criteria use, researchers have rarely explored the feasibility of visualizing and measuring the similarity of relevance judgments based on the use of relevance criteria during the search process. In a study by [4], the level of use and complexity of relevance criteria was plotted visually using different types of colors. In this manner, the density of relevance criteria that tended to be used frequently by users while searching the web could be represented in the form of a heat map. The study presented in this paper made use of Latent Semantic Indexing as a means of representing relevance criteria in a 2-dimensional vector space and computing the level of similarity between judgments using traditional cosine similarity measure from Vector Space Modeling. In particular, the objectives of this study were:

- To show graphically how relevance judgments deviated from the qualitative criteria and characteristics of a user's information need during the search process;
- To compute the level of similarity between relevance judgments, given a specific information need.

The representation of relevance judgments can advance our understanding of human interaction with information on the web. However, it could be also used as an analytical tool for researchers and scientists in information seeking and retrieval research which could assist them make methodological decisions about their data. For example, researchers could use this type of data to identify interesting patterns and relationships between relevance judgments that could be subjected to further examination and experimentation. Also, they could focus their data analysis on a subset of relevance judgments or identify and study abnormalities in the relevance judgment behavior of specific participants. Therefore, as well as to the advancement of our theoretical understanding of human relevance judgment behavior, this type of research could also have methodological implications for researchers.

2. LATENT SEMANTIC INDEXING
Latent Semantic Indexing (LSI) is a mathematical method used in Information Retrieval (IR) to rank documents taking into account both the frequency with which a term *i* occurs in a document *d*, and the underlying semantic association between terms across documents. For example, if a web searcher used the term *actor* in a query she would expect to retrieve documents about *actors* and *acting*. However, to a lesser extent she could be also interested in documents about theaters and/or cinemas. LSI provides the opportunity to identify patterns of semantic relationship between terms, thus, retrieving documents that partially match or are close enough to a user's query by studying the frequency with which specific terms co-occur in documents.

IIIX'12 Nijmegen, The Netherlands.

To perform LSI a $A=(\alpha_{ij})$ matrix of *terms by documents* is constructed, where α_{ij} is the frequency with which term *i* occurs in document *j* (Table 1).

Table 1. Matrix of terms by documents

	Doc 1	Doc 2	Doc 3	Doc n
Term 1	2	4	6	…
Term 2	5	5	4	…
Term 3	0	2	1	…
Term n	…	…	…	…

In order to identify the underlying structure of terms within documents the Singular Value Decomposition (SVD) is applied that decomposes matrix *A* into the product of three matrices: $A=U\,S\,V^T$, where *U* is a term matrix whose columns define term eigenvectors, V_T is a matrix the columns of which define document eigenvectors, and finally *S* is a diagonal matrix where the singular values of *A* are presented diagonally in decreasing order.

Because the eigenvectors presented in matrices *U*, V_T and *S* represent too many dimensions of the relationship between terms and documents, a truncated SVD method is applied. Truncated *U S* and V_T matrices lower the complexity of multi-dimensionality by retaining only the largest *k* vectors and removing the remaining ones from the analysis. The optimal number of retained *k* vectors is debated and it depends on the complexity and heterogeneity of eigenvector values within the matrices. For the purpose of illustration, Figure 1 shows an example of truncated U, S and V_T matrices in a k=2-dimentional space. The values in U and V_T matrices can be used as co-ordinates (x,y) to plot graphically terms and documents respectively. In particular, each row in the U matrix represents the *x* and *y* co-ordinates for each individual term (*i*), while each column of the V_T matrix represents vectors of documents (*j*) of a hypothetical A_{ij} matrix (e.g. V_T = d_1(x-axis): 0.233; (y-axis): 0.343, d_2(x-axis): -0.056; (y-axis): 0.123, d_n…). It is worth mentioning that vectors can be either negative or positive.

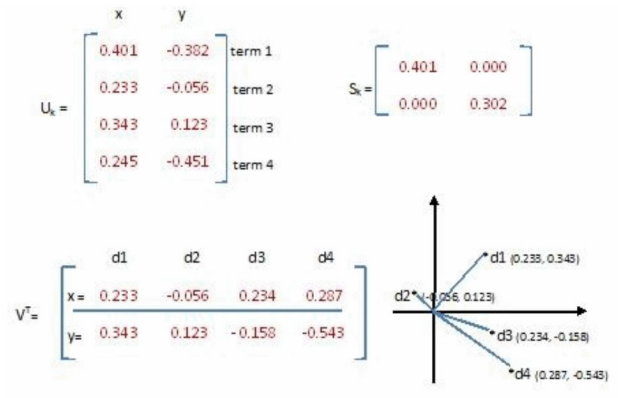

Figure 1. Example of truncated SVD

For the purpose of information retrieval a user's query is represented as a vector in a *k*-dimensional space and compared to document vectors. In order to compute query vectors the following equation can be used:

$$q=q^T U_k S_k^{-1} \qquad (1)$$

Where $q^T U_k$ is the query vector multiplied by the corresponding term vectors in k-dimensional space of the truncated U matrix. The product of this multiplication is then weighted by the truncated diagonal A singular values of the S_k^{-1} matrix. It is worth mentioning that a query vector can be represented in the k-dimensional space. For example, in the 2-dimensional space a query vector consists of two x,y co-ordinates. This happens because query vectors are computed twice for the corresponding x and y term co-ordinates of the truncated U matrix. Finally, having computed query vectors, documents can be ranked against queries using the following cosine similarity measure from the Vector Space Model:

$$\cos(\theta)=\frac{q\cdot d}{|q||d|} \qquad (2)$$

3. METHODOLOGY

LSI was applied to a set of data collected from a user study that investigated the relevance judgment behavior of 24 participants on the web. All participants searched for a real information need using the web, but the analysis of the results was focused on the set of relevance judgments made in the search result interface of the Google search engine. The recording and identification of participants' relevance judgments was based on the use of eye tracking, talk aloud and post-search interviews. Due to length limitations a detailed description of the methodology and findings of the user study are beyond the scope of this paper. Details about the user study can be found in [3]. However, it is important to mention that one of the outcomes of the user study was the development of a matrix which associated relevance criteria with surrogates of the search result interface where these criteria occurred, as well as the corresponding search stage (e.g. for the purpose of data analysis each query submitted by a participant in the Google search engine was considered to be a different search) (Table 2). Since relevance criteria can only implicitly be inferred from the analysis of users' relevance judgment behavior, in the context of this study the identification of relevance criteria and their association with the surrogates and the corresponding search stage where these occurred came from the analysis of transcribed talk aloud protocols and participants' eye movements using a step-wised content analytic technique [3].

3.1 The Data Set

The data set used in the context of the present study was based on the aforementioned matrix that associated relevance criteria with judged surrogates from the search result interface of Google and the corresponding search stage where judgments of surrogates occurred. Table 2 shows an example of this criteria x surrogate matrix. Each cell in the matrix contained values that represented the frequency of occurrence of each criterion within surrogates and across searches per participant. Columns in Table 2 represent individual relevance judgments per ranked surrogate, while rows display the level of use of each criterion across surrogates. In summary, each judged surrogate was considered to be an individual relevance judgment that consisted of one or more relevance criteria that occurred during a specific search stage.

3.2 Applying LSI to Relevance Judgments

LSI was applied to a criterion x surrogate matrix and it was repeated for each participant (Table 2). However, in this case, terms (*i)* were substituted by types of relevance criteria (c), and documents (*d)* by surrogates (*s)*.

Table 2. Matrix sample of relevance criteria x surrogates

Participant id = 01	1st Search			
Criteria	**1st Surrogate**	**2nd Surrogate**	**5th Surrogate**	**n…**
Topicality	2	0	1	…
Scope	1	3	0	…
Quality	0	1	0	…
Recency	2	1	4	…
Criterion n	…	…	…	…

Due to the moderate size of the produced k-dimensional U,S and V_T matrices, the truncated A_k matrices were set to contain the eigenvectors presented in the first two rows of the V_T matrix (which contained surrogate vectors), the vectors presented in the first two columns of the U matrix (these were the vectors for the relevance criteria) and finally, the diagonal eigenvectors of the first two columns of the S matrix. This type of truncation permitted a simplified 2-dimensional representation and comparison of relevance judgments.

3.3 Computing Vectors for Information Needs

Instead of computing query vectors, for the purpose of this study, vectors were computed for the broader information need. This decision was made because the contents of the query are monotonically focused on *Topicality* and therefore a query cannot be compared to or regarded as representative of a users' relevance judgment behavior which is multi-dimensional. As opposed to standard queries, information needs can expose a richer context of criteria that people tend to use when they judge relevance.

Data about participants' information needs were collected during the user study. In particular, each participant was asked to articulate her information need before searching for information on the web:

> *I am looking for information* ***about Easter Islands****. I need to find* ***articles or reports or visual material like pictures and videos*** *about rock art in the islands. I am particularly* ***focused on the relationship*** *between rock art and local mythology* (Participant id = 05)

The boldfaced words and phrases in the above example represent the presence of relevance criteria that probably the participant would use in order to judge relevance of information. For example, a total of three criteria can be identified for the above example. These are: **Topicality** (…*about Easter Islands*…); **Resource type** (…*articles or reports…videos*…); and **Scope/Specificity** (…*focused on the relationship*…*local mythology*). Therefore, the information need was decomposed into its constituent relevance criteria. Having identified a set of criteria the vectors for information needs were computed following equation (1). However, in this case terms (t) in queries (*q)* were substituted by relevance criteria (c) in Information Needs (*I*):

$$I = I^T U_k S_k^{-1} \quad (3)$$

Finally, judgments made in surrogates of the Google search result interface can be ranked against the vectors of the information need using the cosine similarity measure in equation (2). Like in the case of equation 1, query vectors are substituted by the vectors of the information need (*I*) while the document vectors (*d)* are replaced by the surrogate vectors (*s*):

$$\cos(\theta) = \frac{I \cdot s}{|I||s|} \quad (4)$$

4. RESULTS

To illustrate the proposed technique we focus on the interpretation of data generated for one participant. This decision was made due to the short size of this paper as well as because it would be more interesting to explain how researchers can use the outputs of LSI in order to interpret and process relevance judgment behavior data at the individual user level. Also, we decided not to compare judgments made between participants. Since participants in the study had different information needs it was impossible to use a global cosine similarity measure in order to compare the relevance judgments made by user A against the vectors produced for the information need of another user B. Therefore, at this stage of the analysis it is more appropriate to focus on the results obtained from a single user's interaction with the search result interface of the Google search engine across a set of successive searches.

Figure 2 and the corresponding Table 3 present the judgment behavior of a participant who searched for information about *Bret Easton Ellis' novel American Psycho*. Specifically, the participant was interested in finding more about the theatrical adaptation of this work rather than about the movie or the book. In this manner, two main relevance criteria were assigned to the participant's information need. These were: **Topicality** (*Bret Easton Ellis' novel American psycho*) and **Scope/Specificity** (*...keen to learn more about the theatrical adaptation of his work...*). The participant performed a set of five successive searches in the Google search engine in order to find relevant information. Table 3 represents the level of cosine similarity between the relevance criteria used to judge the relevance of each surrogate (s1, s2, s3….s_n) and the criteria identified in the actual Information Need (IN) across the five searches. Figure 2 shows an example of the graphical representation of the relevance judgments made for each surrogate in the case of the first search only.

For any judged surrogate in Figure 2, a dissimilarity between the criteria of the information need (*IN*) and those of the surrogate is represented graphically by a larger distance in the Euclidean 2-dimensional vector space between a point S_n and the *IN* vector co-ordinates. The positioning of an S_n or *IN* point in the 2-dimensional space is based on the (x,y) co-ordinates of the V_T matrix – for surrogates (*s)*– and the corresponding output of equation (3) – for the information need (IN). In particular, Figure 2 graphically presents the level of similarity between the criteria identified in the participant's *IN* and the criteria used to judge the relevance of information displayed in seven surrogates (s1, s2, s3, s4, s5, s6 and s7 – i.e. these were the top seven surrogates of the first page of the search result interface). Specifically, the figure shows that there was an exact match between the criteria used to judge relevance of information in the first surrogate (*s1*) and the *IN*. From the data displayed in the original *criteria x surrogate* matrix (that recorded data about the type and frequency of

occurrence of relevance criteria across surrogates) we know that these were the criteria of *Scope* and *Topicality*. The exact match between the eigenvectors of both points *s1* and *IN* in figure 2 showed consistency between the participant's judgment and the actual information need. Alternatively, the exact match between *s1* and *IN* is represented also by the ranking scores of the cosine similarity measure in Table 3. Surrogate *s1* has a score of 1.0 for the first search. This score denotes the presence of an exact match between the judged surrogate and the *IN*. It is worth mentioning that cosine similarity rankings can range between 1.0 (exact similarity) and -0.1 (exact dissimilarity).

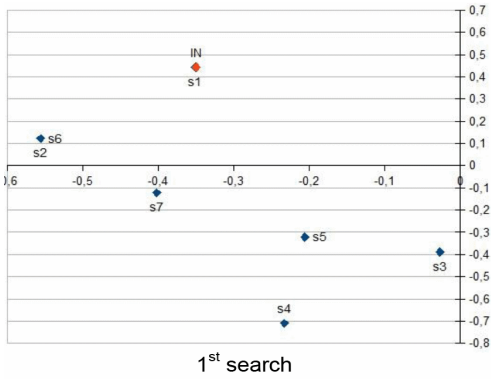

Figure 2. Information Need and Judgment vectors

Also, Figure 2 shows that judgments of relevance deviated from the initial information need (*IN)* during the evaluation of the remaining surrogates of the first search session and this is how thegraphical representation in Figure 2 and the scores in Table 3 help us identify where different judgments of relevance are being used. For instance, Table 3 shows that judgments in surrogates (*s2* and *s6*) were ranked second in terms of similarity with the *IN* vectors. This deviation in similarity can be observed graphically in Figure 2.

Table 3. Ranking of relevance judgments across searches

1st Search	2nd Search	3rd Search	4th Search	5th Search
s1 = 1.0	s2 = 1.0	s1 = 1.0	s1 = 0.9	s8 = 0.9
s2 = 0.8	s1 = 0.95	s9 = 0.95	s4 = 0.9	s9 = 0.8
s6 = 0.8	s3 = 0.95	s2 = 0.9	s2 = -0.4	s3 = 0.65
s7 = 0.4	s4 = 0.9	s3 = 0.85	s3 = -0.7	s10= 0.6
s5 = -0.3	s5 = 0.5	s6 = -0.1		s1 = 0.6
s4 = -0.5	s7 = -0.1	s10 = - 0.5		s2 = 0.5
s3 = -0.7	s6 = -0.4			s7 = 0.3
				s5 = 0.25
				s6 = 0.25
				s4 = 0.2

A closer examination of the data recorded in the original criteria x surrogate matrix showed that this happened because as well as to the criteria of Scope and Topicality the participant also used the criterion of Quality to judge relevance of the information presented in surrogates ranked in positions two and six of the search result interface (s2 and s6). Quality was not a criterion identified in the initial information need. Similarly, in the case of the surrogate ranked in position seven (s7), the participant used criteria that partially matched those defined by the information need. These were the criteria of Topicality (which matched exactly with the criterion of Topicality of the corresponding information need IN) and Quality (lack of match with the information need IN). The partial dissimilarity in judgments between the IN and the surrogates s2, s6 and s7 in the case of the first search stage is shown in Table 3. This Table ranks judgments made across surrogates according to their cosine similarity with the IN.

The largest dissimilarity or anti-similarity was observed for judgments made in surrogates ranked in positions three, four and five (s3, s4 and s5). This happened because the participant made judgments without using any of the criteria identified in the information need. For example, in the case of surrogates s5 and s3 the participant used respectively the criteria of Quality and Tangibility only, while both criteria were used for judgments made in surrogate s4. Judgments made on these surrogates scored negative cosine similarity score and their deviation from the IN vector co-ordinates is denoted by a large distance in the Euclidean 2-dimensional space in Figure 2.

4.1 Co-occurrence of Relevance Criteria

Although in the case of surrogates *s5*, *s4* and *s3* of the first search the participant did not use any of the criteria identified in the corresponding information need (*IN*), the judgment made in surrogate *s5* was ranked higher in terms of similarity than judgments made in surrogates *s4* and *s3*. This happened because LSI valued the co-occurrence of relevance criteria across relevance judgments. For example, as opposed to the criterion of *Tangibility*, the criterion of *Quality* tended to co-occur more frequently in relevance judgments with the criteria of *Scope* and *Topicality* (both present in the information need). Therefore, surrogates with judgments related to the criterion of *Quality* were ranked higher in terms of similarity with the information need than surrogates for which judgments were related to the criterion of *Tangibility*. LSI is effectively picking up latent relevance criteria – ones not expressed in the original need but implicit to this need (i.e. they are implicit because they are not explicitly stated but are often used in assessment).

5. CONCLUSIONS

The results showed that the dissimilarities between relevance judgments and the underlying information need can be measured quantitatively through the use of LSI. These dissimilarities are indicative of cognitive shifts and changes in users' judgment strategies, but should not be interpreted as the result of a good or bad judgment. Thus, the use of LSI for the analysis of relevance judgments should not be confused with its role in traditional IR research (e.g. TREC and INEX experiments).

6. REFERENCES

[1] Tombros, A., Ruthven, I. and Jose, J. (2005). How users assess web pages for information seeking. J. Am. Soc. for Information Science & Technology, 56(4), 327-344.

[2] Spink, A. and Greisdorf, H (2001). Regions and levels: measuring and mapping users' relevance judgments. J. Am. Soc. for Information Science & Technology,52(2), 161- 173.

[3] Balatsoukas, P. and Ruthven, I. (2010). The use of relevance criteria during predictive judgment: an eye tracking approach. In: 73rd Annual Meeting of the American Society for Information Science & Technology, 47, 1-10.

[4] Beresi, U., Kim, Y., Song, D. and Ruthven, I. (2010). Why did you pick that? Visualizing relevance criteria in exploratory search. Int. J. Digit Lib, 11, 59-74.