Integrating Similarity Measures in Internet Resource Discovery

Adanki Padma,Conception L Khan, K G George, Francis Jawahar Devadason,
R Sadananda

Computer Science and Information Management Program
School of Advanced Technologies
Asian Institute of Technology
Thailand
Sada@cs.ait.ac.th

Abstract. The promise of the information age entails controlling the potentially unwanted flux of information from sources to their receiving end-users and delivers just the relevant information. Moreover the Internet size necessitates scalability of algorithms used to locate and track resources. As the number of these Internet servers increases rapidly, it becomes difficult to determine the relevant servers when searching for information. To address this issue, it is necessary to formalize the queries and the servers and develop a measure of similarity between them. This paper reviews various measures of similarity and proposes a new measure, which is an improvement over the existing measures. The focus is on Boolean environments, as it is generally believed that Boolean expressions can precisely describe a server's contents as well as a user's information need. Users can send queries to the top-ranked systems and obtain most of the relevant information. To compare the performance of the existing measure, we conducted two case studies on Compendex databases. The paper includes the details of the case study experiments and performance comparisons of the proposed measure with those available. The results are found to be encouraging.

1. Introduction

The servers on the Web represent a vast potential resource. For novice users, they have no idea where to send requests. For experienced users, they may miss new servers having information they want. Because each users requests vary, it is inappropriate to broadcast requests to all servers. This overwhelms the underlying networks and overloads irrelevant servers. One step to solve this problem is to determine relevant servers before the real search starts.

The emerging suite of tools for finding and discovery of interesting information from the internet such as WWW, Gopher, Archie, Nomenclator, WAIS etc., lack some functionality's in user's interest mapping and sometimes return inconsistent information. In the current scenario, and anticipating the importance of this technology for future research we have pursued the study on the similarity measures that can greatly improve the searching process in today's world of overwhelming information. Boolean similarity measures gives users a ranked list of relevant sources based on their similarities with respect to a query so that it becomes easy to search and locate desired information and filter-out unwanted ones.

The main objective of this paper is to introduce a new and efficient Boolean Similarity measure that is based on Jaccard Coefficient, and is an improvement over an existing method and compare the performance of the new measure and the existing measure against the Jaccard Coefficient (applied on an operational information retrieval system (IRS)) and present the experimental results. In section II we discuss the existing similarity measures and in section III we introduce a new Boolean similarity measure. In section IV, the case studies and the methodology for comparing the performance of the existing similarity measures with the new similarity measure as well as the implementation of the new similarity measure using client-directory-server model is discussed. Finally the conclusions are discussed in section V.

2. Existing Similarity Measures

The degree of similarity as determined by the document retrieval system, between Document Representations (DR) and Search Request Formulations (SRF) is the system's estimation of the likelihood of usefulness (relevance) of a document with respect to the user's query. It can be well explained by the Jaccards coefficient which is one of the popular coefficients of associations [Rijsbergen, 1975].

Let X and Y be the set of index terms occurring in two document (or request) representatives.

Then according to Jaccard's Coefficient

$$\frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

$|X \cap Y|$ gives the number of shared index terms.

$|X \cup Y|$ accounts for all the information contained in X and Y

An adaptation of the Jaccard's Coefficient to a given information retrieval system as proposed by [Radecki, 1982], lead to the formulation of the query similarity measure S as given below.

Let $\psi(Q)$ be the response sets of documents applied on a document representation R, containing $\varphi(R)$ set of documents. The similarity value S between Q and R is then defined as the ratio of the number of common documents to the total number of documents in $\psi(Q)$ and $\varphi(R)$.

$$S(Q, R) = \frac{|\psi(Q) \cap \varphi(R)|}{|\psi(Q) \cup \varphi(R)|} \tag{2}$$

Where $\cap$ denotes set intersection, $\cup$ denotes set union and $|.|$ gives the size of a set.

Since all the documents satisfying query Q belong to cluster R (i.e., $\psi(Q) \subseteq \varphi(R)$),

$$S(Q, R) = \frac{|\psi(Q)|}{|\varphi(R)|} \tag{3}$$

The similarity measure based on Jaccard coefficient as presented above is one of the accurate measures and may be used in all types of IRSs. It may work well in traditional IRSs, which has documents in a single database or where the database size is small. But in a distributed environment such as in Internet, to determine the relevant servers based on the estimation of "similarity", a query has to be sent to each one of the information servers to determine the coefficient of association and then rank the servers relevant to the user's request in a descending order. Furthermore, if the similarity is calculated based on the query results from every information system, the searching order is no longer needed because already the system has been searched.

In order to overcome the limitation mentioned above, [Radecki, 1985] proposed another similarity measure S* that is based on only the to-be-compared Boolean expressions which are Boolean combinations of index terms and is independent of the responses to the queries and hence independent from the variation in the subject matter of document collections. According to this similarity measure, the system response to a given query is the set of documents whose representations satisfy the logical requirements of the associated Boolean SRF. Any Boolean SRF can be represented in disjunctive normal form, which is a union of set of documents whose representations are true with respect to the corresponding index terms.

Let q_i and q_j be two Boolean SRFs consisting of $T_{i,}T_j$ of index terms. In S* each Boolean expression is transformed into its Reduced Disjunctive Normal Form (RDNF), which is the disjunction of a list of reduced atomic descriptors. This is done in the following steps:

Distribution: Distribution is performed on the query q_i by applying the distribution law, so as to transform it into its CDNF (Compact Disjunctive Normal Form) which is a disjunction list of Compact Atomic Descriptors (CAD) as given below:

$$(t_1 \vee t_2) \wedge t_3 = (t_1 \wedge t_3) \vee (t_2 \wedge t_3), \tag{4}$$

Where t_1, t_2 and t_3 are descriptors and $(t_1 \wedge t_3)$, $(t_2 \wedge t_3)$ are compact atomic descriptors

Transform: Each compact atomic descriptor is expanded to contain all the index terms in both the queries put together, i.e., ($T_i \cup T_j$ of descriptors) in both its original and negated form. This can be done by multiplying those index terms from T that are not present in the compact atomic descriptor using:

$$t_a = (t_a \wedge t_b) \vee (t_a \wedge \sim t_b) \tag{5}$$

Where t_a is a compact atomic descriptor and t_b is the index term that is not present in t_a. The resultant is a reduced atomic descriptor that contains all the descriptors (original or negated) in the to-be-compared Boolean expressions. The disjunction list of such reduced atomic descriptors forms an RDNF. Thus, S* proposed by Radecki, using Jaccard's coefficient between the two RDNF's is defined as the ratio of the number of common reduced atomic descriptors in q_i and q_j to the total number of reduced atomic descriptors in them as given below:

$$S^*(q_i, q_j) = \frac{|(q_i)_{T_i \cup T_j} \cap (q_j)_{T_i \cup T_j}|}{|(q_i)_{T_i \cup T_j} \cup (q_j)_{T_i \cup T_j}|} \tag{6}$$

Where $(q_i)_{T_i \cup T_j}$ and $(q_j)_{T_i \cup T_j}$ are the **RDNFs** (Reduced Disjunctive Normal Form) of q_i and q_j.

Example 1: Let, $q_1 = (t_1 \wedge t_2) \vee {\sim}t_5$
$q_2 = (t_2 \vee t_3) \wedge t_5$

In this case, $T_1 = \{t_1, t_2, t_5\}$ and $T_2 = \{t_2, t_3, t_5\}$. Hence $T_1 \cup T_2 = \{t_1, t_2, t_3, t_5\}$

Step 1: To calculate S*, we first perform distribution on the queries (q_1) and (q_2) to obtain their compact atomic descriptors

$$(q_1) = t_1t_2 \vee {\sim}t_5$$
$$(q_2) = t_2t_5 \vee t_3t_5$$

Step 2: Expand each compact atomic descriptor of each query to obtain their RDNFs $(q_1)T_1 \cup T_2$ and $(q_2)T_1 \cup T_2$ respectively.

$$(q_1)T_1 = t_1t_2t_5 \vee t_1t_2{\sim}t_5 \vee t_1{\sim}t_2{\sim}t_5 \vee {\sim}t_1t_2{\sim}t_5 \vee {\sim}t_1{\sim}t_2{\sim}t_5$$
$$(q_1)T_1 \cup T_2 = t_1t_2t_3t_5 \vee t_1t_2{\sim}t_3t_5 \vee t_1t_2t_3{\sim}t_5 \vee t_1t_2{\sim}t_3{\sim}t_5 \vee t_1{\sim}t_2t_3{\sim}t_5 \vee t_1{\sim}t_2{\sim}t_3{\sim}t_5 \vee$$
$${\sim}t_1t_2t_3{\sim}t_5 \vee {\sim}t_1t_2{\sim}t_3{\sim}t_5 \vee {\sim}t_1{\sim}t_2t_3{\sim}t_5 \vee {\sim}t_1{\sim}t_2{\sim}t_3{\sim}t_5$$
$$(q_2)T_2 = t_2t_3t_5 \vee t_2{\sim}t_3t_5 \vee {\sim}t_2t_3t_5$$
$$(q_2)T_1 \cup T_2 = t_1t_2t_3t_5 \vee {\sim}t_1t_2t_3t_5 \vee t_1t_2{\sim}t_3t_5 \vee {\sim}t_1t_2{\sim}t_3t_5 \vee t_1{\sim}t_2t_3t_5 \vee {\sim}t_1{\sim}t_2t_3t_5$$

Thus, $S^*(q_1, q_2) = \frac{2}{14} = 0.143$

Applying the above example to our Query-Server environment, if a user query is compared against N server descriptions, it needs 2N RDNF transformations to calculate the similarity between them. This method suffers when the number of server descriptions is large and users query frequently. The system will spend significant amount of time recomputing RDNFs and consequently will perform badly. To solve this problem, a new similarity measure is suggested in this paper which is a modification of Radecki's method, based on the idea of Shih-Hao Li (1997), so that it need not recompute RDNFs of server descriptions while still providing statistically equivalent results.

3. New Similarity Measure

The new similarity measure $S^{\ominus}$ is an improvement over the Radecki's RDNF approach which is independent of the underlying information systems, independent from the variation in the subject matter of document collections as well as requires less computation, space and time. Moreover, it is based on only the descriptors in its queries and not introducing new descriptors from the other query pair i.e., it is independent of those in the other, to-be-compared query pair.

To be specific, the new measure is dependent only on the CDNFs i.e., on the results of the step 1 alone obtained in Radecki proposed S*, that consists of only a subset of descriptors occurring in its own disjunctive normal form. This can be explained further.

Let Q and R represent two Boolean expressions and TQ and TR the sets of the descriptors that appear in Q and R respectively. $\overline{Q}$ and $\overline{R}$ are the CDNFs of Q and R respectively. $\overline{Q}^i$ indicates the ith compact atomic descriptor of $\overline{Q}$, $\overline{R}^j$ the jth compact atomic descriptor of $\overline{R}$, and T_Q^i and T_R^j the sets of descriptors in $\overline{Q}^i$ and $\overline{R}^j$, respectively. Suppose there are k common descriptors between T_Q^i and T_R^j, i.e., $| T_Q^i \cap T_R^j | = k$, and m and n are the numbers of remaining (unique) descriptors in T_Q^i and T_R^j respectively. Then

$$\overline{Q}^i = (c_1 \wedge c_2 \wedge \ldots c_k \wedge a_1 \wedge a_2 \wedge \ldots \wedge a_m) \qquad (7)$$

$$\overline{R}^j = (c_1 \wedge c_2 \wedge \ldots c_k \wedge b_1 \wedge b_2 \wedge \ldots \wedge b_n) \qquad (8)$$

$$T_Q^i = \{c_1, c_2, \ldots, c_k, a_1, a_2, \ldots, a_m\} \qquad (9)$$

$$T_R^j = \{c_1, c_2, \ldots, c_k, b_1, b_2, \ldots, b_n\} \qquad (10)$$

Where $c_1 \ldots c_k$ are the common descriptors between $\overline{Q}^i$ and $\overline{R}^j$, $a_1 \ldots a_m$ and $b_1 \ldots b_n$ are their remaining descriptors. From observing the above it is obvious that the numbers of reduced atomic descriptors in $(\overline{Q}^i)_{T_Q^i \cup T_R^j}$ is 2^n and in $(\overline{R}^j)_{T_Q^i \cup T_R^j}$ it is 2^m respectively. Applying eqn. 6,

Since $|A \cup B| = |A| + |B| - |A \cap B|$, from Equation 14 it follows that

$$S^*(\overline{Q}^i, \overline{R}^j) = \frac{|(\overline{Q}^i)_{T_Q^i \cup T_R^j} \cap (\overline{R}^j)_{T_Q^i \cup T_R^j}|}{|(\overline{Q}^i)_{T_Q^i \cup T_R^j}| + |(\overline{R}^j)_{T_Q^i \cup T_R^j}| - |(\overline{Q}^i)_{T_Q^i \cup T_R^j} \cap (\overline{R}^j)_{T_Q^i \cup T_R^j}|} \qquad (11)$$

$$\frac{1}{2^n + 2^m - 1} \qquad (12)$$

$$\underline{1} \qquad \textbf{(13)}$$

$$2^{|T_R^j - T_Q^i|} + 2^{|T_Q^i - T_R^j|} - 1$$

Similarly, if a reduced atomic descriptor in $(\overline{Q}^i)_{T_Q^i \cup T_R^j}$ contains t and every reduced atomic descriptor in $(\overline{R}^j)_{T_Q^i \cup T_R^j}$ contains ~t, then

$$S^*(\overline{Q}^i, \overline{R}^j) = 0 \qquad \textbf{(14)}$$

The individual similarity measure s^Θ may then be defined as
$s^\Theta(\overline{Q}^i, \overline{R}^j) = 0$ if $T_Q^i \cap T_R^j = \phi$ or $\exists t$ such that $t \in T_Q^i$, $\sim t \in T_R^j$, otherwise it is

$$= \frac{1}{2^{|T_R^j - T_Q^i|} + 2^{|T_Q^i - T_R^j|} - 1} \qquad \textbf{(15)}$$

where $|T_R^j$ - $T_Q^i|$ is the number of descriptors that appear in T_R^j but not in T_Q^i and $|T_Q^i$ - $T_R^j|$ is the number of descriptors that appear in T_Q^i but not in T_R^j.
The similarity measure S^Θ between Q and R is defined as the sum of the individual s^Θ given by

$$S^\Theta(Q,R) = \sum_{i=1}^{|Q|} \sum_{j=1}^{|R|} s^\Theta(\overline{Q}^i, \overline{R}^j) \qquad \textbf{(16)}$$

Where $|Q^i|$ and $|R^j|$are the number of compact atomic descriptors in $\overline{Q}$ and $\overline{R}$ respectively.
As shown above, the individual similarity measure can be obtained from examining the difference between the two CDNFs without transforming to their RDNFs. Hence it avoids the complicated computing process of RDNFs.
Example 2: Following the same details as given in example 2,

$q_1 = (t_1 \wedge t_2) \vee \sim t_5$
$q_2 = (t_2 \vee t_3) \wedge t_5$
On applying distribution, the compact atomic descriptors of q_1 and q_2 are,
$(\overline{q}_1) = t_1t_2 \vee \sim t_5$
$(\overline{q}_2) = t_2t_5 \vee t_3t_5$

$\overline{q}_1^1 = t_1t_2$	$T_{q_1}^1 = \{t_1, t_2\}$
$\overline{q}_1^2 = \sim t_5$	$T_{q_1}^2 = \{t_5\}$
$\overline{q}_2^1 = t_2t_5$	$T_{q_2}^1 = \{t_2, t_5\}$
$\overline{q}_2^2 = t_3t_5$	$T_{q_2}^2 = \{t_3, t_5\}$

Where $T_{q_1}^1, T_{q_1}^2$, $T_{q_2}^1$ and $T_{q_2}^2$ are the sets of descriptors in the compact atomic descriptors in $\overline{q}_1^1, \overline{q}_1^2, \overline{q}_2^1$ and $\overline{q}_2^2$ respectively. The individual similarity measures are therefore

$$s^{\Theta}(\overline{q}_1^1, \overline{q}_2^1) = \frac{1}{2^1 + 2^1 - 1} = 0.333$$

$$s^{\Theta}(\overline{q}_1^1, \overline{q}_2^2) = 0$$

$$s^{\Theta}(\overline{q}_1^2, \overline{q}_2^1) = \frac{1}{2^0 + 2^1 - 1} = 0.5$$

$$s^{\Theta}(\overline{q}_1^2, \overline{q}_2^2) = \frac{1}{2^0 + 2^1 - 1} = 0.5$$

$$S^{\Theta}(q_1, q_2) = s^{\Theta}(\overline{q}_1^1, \overline{q}_2^1) + s^{\Theta}(\overline{q}_1^1, \overline{q}_2^2) + s^{\Theta}(\overline{q}_1^2, \overline{q}_2^1) + s^{\Theta}(\overline{q}_1^2, \overline{q}_2^2) = 1.333$$

Also the similarity values calculated using S^{Θ} in example 3 are different from those calculated using S* in example 2. S* and S^{Θ} are measured on relative scales and hence cannot be compared directly. However, they can be used to rank a list of Boolean expressions measured by the same method. Accordingly, we used the rankings obtained by the similarity measures S* and S+ and carried out two case studies to compare the rankings estimated by these similarity measures.

4. Case Study Experiments

Two case studies were undertaken and analyzed to compare the performance of the existing similarity measure S* and the new similarity measure S^{Θ}. The standardized and widely used "Compendex" data that provides coverage of the world's significant engineering and technical literature in various technical disciplines, forms the datasets for the current research. Please refer to Table 1 and 2. For the case study I, the Compendex dataset for the period of January to September 1997 is used for the purpose of the study. A set of 12 Boolean search request queries have been formulated using five descriptors (keywords), with an average of 3 terms/query. Similarly, for the case study II, the Compendex dataset for the period of January to September 1994 is utilized. A set of 11 Boolean search request queries have been formulated using five descriptors (keywords), with an average of 3.4 terms/query.

4.1 Methodology

The aim of the case studies - is to find out how strong the new similarity measure which is based on CDNF is, when compared to its original method based on RDNF. Thus to compare the old method and the new method, a benchmark similarity measure - an application of Jaccard Coefficient is chosen against which the estimations obtained from two methods can be contrasted. In other words, the actual responses from a real (operational) information system can be the closest criterion to justify the two measures in consideration.

Benchmark measure - We were convinced of using the similarity measure S (Equation 1) that is based on Jaccard Coefficient. The justification for using such similarity

measures for queries is based on the expectation that the similarity value between queries is associated strongly with the degree of overlap of the responses to the queries. Secondly, the advantage of this method for determining the similarity values between queries is that it may be applied in every information retrieval system with unordered responses, regardless of the way in which the documents are represented and queries are formed.
The steps for comparing the performance of the similarity measures under study are as follows:

1. To apply Jaccard coefficient on the Compendex databases; for a given case study, each of the N queries is applied on the database and the responses to the queries is obtained to create individual servers, where each server description is represented by the query which acts as the filter. Next, each query (N-1 queries) is submitted to each of the N servers. The number of hit documents is used to calculate S using Equation 3. Based on the S values from the N servers, we then rank them for each query and use that as the standard ranking to evaluate S^* and $S^{\ominus}$.

2. We applied (6) and (16) to each filter-query pair (i.e., query pair) and calculated the similarity values between the queries for each pair $S^*(q_i, q_j)$ and $S^{\ominus}(q_i, q_j)$ where i, j = 1, 2, ...N, as determined by the measures tested respectively.

3. We first wanted to determine whether the three set of rankings are related to each other. This could be done in two steps:
a) Using Kendall's coefficient of concordance, W which expresses the degree of association among clusters of variables. In other words, When we have k sets of rankings, we may determine the association among them or the divergence of the actual agreement between k set of rankings from the maximum possible (perfect) agreement (Siegel, 1956). The general form to compute W is as follows, considering that when tied observations occur, the observations are each assigned the average of the ranks they would have been assigned had no ties occurred, the usual procedure in ranking tied scores.

$$W = \frac{s}{\frac{1}{12}k^2(N^3 - N) - k\sum_T T} \qquad \textbf{(17)}$$

Where s = sum of squares of the observed deviations from the mean of R_j, i.e.,
$s = \sum (R_j - \sum R_j/N)^2$
k = number of sets of rankings
N - number of entities (objects) ranked
$\frac{1}{12}k^2(N^3 - N)$ = maximum possible sum of the squared deviations, i.e., the sum s w which would occur with perfect agreement among k rankings.
The correction factor, $T = \frac{\sum (t^3 - t)}{12}$
Where t = number of observations in a group tied for a given rank
$\sum$ Sum over all groups of ties within any one of the k rankings.
Where $\sum_T T$ is the sum of all the values of T for all the k rankings.

b) Test the significance of the observed value of W by finding chi square χ^2 and then determined the probability associated with the occurrence of the value as large as the s with which it is associated by referring the calculated χ^2 value with that of the χ^2 critical values as given in the table C of the Appendix (Siegel, 1956).

$$\chi^2 = k(N-1)W \qquad \textbf{(18)}$$

During our research on both the case studies, we found that there exists strong agreement between the values obtained by the three similarity measures. This indicates a very high probability associated with the observed value of W that enables to accept the null hypothesis that the similarity measures between the three sets of rankings S, S* and $S^{\ominus}$ are related to each other. Hence we proceeded with the following step to find out which of the two measures - the existing S* or the new $S^{\ominus}$, generates a ranking closer to the Jaccard coefficient S.

4. Spearman's Rank Order Correlation Coefficient (r_s) is a non-parametric measure that has been used in this research to compute the degree of association between (S*,S) and between ($S^{\ominus}$,S). The important feature of this coefficient is that it does not require any assumptions to be made about the shape of the population from which the scores were drawn. The only requirement underlying the use of this coefficient is that both variables under study should be measured in at least on an ordinal scale, so that the scores corresponding to the variables may be ranked in two ordered series. The tested similarity measures do meet this requirement, so there are no arguments against using r_s to test our research hypothesis. The general form of the formula in question is as follows:

$$r_s = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma \delta^2}{\sqrt{2\Sigma x^2 \ \Sigma y^2}} \qquad \textbf{(19)}$$

Where $\Sigma x^2 = \frac{N^3 - N}{12} - \Sigma T_x$ and

$$\Sigma y^2 = \frac{N^3 - N}{12} - \Sigma T_y$$

In the above two equations, T is the correction factor similar to that of Kendall's coefficient of concordance.

Accordingly, T_x and T_y stand for the sums of the values of T_x and T_y for all the various sets of tied scores corresponding to the variables X and Y, respectively.

4.2 Results: Case Study I

The results of the actual hit documents on the compendex system S are presented in Table 3. Following the five steps described previously, we calculated Kendall's Coefficient of Concordance - W and the test of significance - χ^2 which are presented in table 5. According to the values obtained for the data, it is found that for 9 out of 12 cases, the observed value of χ^2 is greater than the table value. This indicates a very high probability associated with the observed value of W that enables to reject the null hypothesis that the similarity measures between the three sets of rankings S, S* and $S^{\ominus}$

are unrelated to each other. In other words, there is a strong association between the new similarity measure and the existing measure.

Hence, the calculations on Spearman rank order correlation coefficient is justified which is helpful to compare which of the two methods generates a ranking closer to the standard, i.e., to compute the degree of association between (S, S*) and (S, $S^{\ominus}$). A summary of the r_s values obtained for all the queries applied against each server in the experiment are presented in table 5. The results indicate that in 8 out of 12 cases, r_s(S, $S^{\ominus}$) is greater than r_s(S, S*). This indicates $S^{\ominus}$ generates a ranking closer to that of S for 8 out of 12 times.

Finally, to consolidate the overall performance of the similarity measures all the observations from 1 to N, where N is 66 in the case study are used and the Spearman's rank correlation coefficient is computed. The correlation coefficient obtained for r_s(S, $S^{\ominus}$) and r_s(S, S*) are 0.704 and 0.470 respectively. This indicates, $S^{\ominus}$ is more stronger than S*. The t test of significance in our case are 7.933 and 4.260 respectively which are highly significant in both the cases at 99% level of significance. It means, for 99 cases out of 100, the new similarity measure is closer to that of the required criterion.

4.3 Results: Case Study II

The results obtained are similar to that of the case study 1. The results obtained from the retrieval system S (actual hits) is given in table 4. The Kendall's Coefficient of Concordance - W and the test of significance - χ^2 are presented in table 5. According to the values obtained for the data, it is found that for 6 of the 11 cases, the observed value of $S^{\ominus}$ is greater than the table value. As in the case of case study I, this again indicates that the similarity measures between the three sets of rankings S, S* and $S^{\ominus}$ are related to each other. The results of Spearman rank order correlation coefficient indicates that in 8 out of 11 cases, r_s(S, $S^{\ominus}$) is greater than r_s(S, S*). The Spearman rank correlation coefficient for all 55 observations put together are 0.550 and 0.419 respectively. The data used are presented in Appendix B2.2. The t test vales obtained are 4.796 and 3.262 respectively which are also highly significant at 99% level of significance. It means, for 99 cases out of 100, the new similarity measure is closer to that of the required criterion.

4.4 Association between the Similarity Measures:

Overall, it has been shown in both the case studies, the results are favorable to the $S^{\ominus}$ measure. The results of Kendall's Coefficient of Concordance and the χ^2 tests indicate that the ranking generated in all the three similarity measures namely, S, S* and $S^{\ominus}$ are the same. Further analysis using Spearman rank order correlation and the t test indicates that the new similarity measure $S^{\ominus}$ is much more closely associated with the criterion (Jaccard coefficient) when compared to that of the existing measure S* which may be imparted to the computational effect of the new similarity measure. In the case study I, the value of r_s ($S^{\ominus}$,S) is greater than r_s(S*,S) for 8 times out of 12 (the x's above zero) and less than r_s(S*,S) for 3 times. Out of these 3 one is very close to the zero border. The results indicate that $S^{\ominus}$ generates a ranking closer to that of S for 67% of the queries. Similarly in the case study II, the value of r_s ($S^{\ominus}$,S) is greater than

$r_s(S^*,S)$ for 8 times out of 11. This indicates that $S^{\ominus}$ generates a ranking closer to that of S for 73% of the queries.

5. Conclusions

In this paper, various approaches to determining similarity values among queries have been examined. More specifically, the investigations reported here have been concerned with the methodology for calculating similarity values among queries that are characterized by Boolean combinations of terms. A new and improved method of similarity measure has been introduced using compact disjunctive normal form (CDNF) to rank the similarity between Boolean expressions that is based on Radecki's Reduced Disjunctive Normal Form (RDNF) of Boolean similarity measure. We compared both the methods i.e., our new method with Radecki's measure on two databases and used the Kendall's Coefficient of Concordance, χ^2 test, Spearman rank coefficients and the t test to show that the new method can get a closer ranking order to that generated by Jaccard's coefficient. The theoretical analysis as well as the case study results prove that this new measure outperforms the one proposed by Radecki significantly in terms of computation, time and space complexity. These results demonstrate that the new similarity measure can greatly improve the searching process in today's world of overwhelming information.

In addition to ranking results, the similarity estimates can be used to help identify similar but autonomously managed retrieval systems. For example, the similarity measure can be used to cluster servers with similar descriptions in a single directory entry. For people using Boolean expressions to represent their interests, such as user profile (Danilowicz, 1994) this similarity measure can help find other individuals having common interests, so that they may share their collections. Our method can also benefit systems that support automatic query formulations by relevance-feedback where the reformed queries could be in complex Boolean forms (Frants and Shapiro, 1991; Salton et al., 1985).

Further improvements to our methodology for determining similarity values among queries could be made by developing procedures for assigning weights to particular query attributes (i.e., reduced atomic descriptors or atomic descriptors), since it is clear that the importance of each individual attribute of a query is not equal (is varied) from the point of view of the information needs of the user. Moreover, it would be advisable to undertake research aimed at determining whether or not any gains could be obtained when relations among descriptors are taken into account.

Table 1: Names of Descriptors used

Descriptors	Case Study I	Case Study II
t_1	TESTING	NETWORK
t_2	SOFTWARE	INFORMATION
t_3	CONTROL	DATA

t_4	COMPUTER	PROTOCOL
t_5	PROCESS	TRANSFER

Table 2: Search Request Formulations

Queries	Case Study I	Case Study II
Q_1	$\sim t_5 \vee (t_1 \wedge t_2)$	$(t_1 \wedge t_2 \wedge t_3) \vee t_4$
Q_2	$\sim t_1 \vee (t_1 \wedge t_3)$	$(t_3 \wedge t_4) \vee (t_2 \wedge t_3)$
Q_3	$\sim t_4 \vee (t_1 \wedge t_2)$	$(t_2 \wedge t_5) \vee (t_2 \wedge t_4)$
Q_4	$(t_2 \vee t_3) \wedge t_5$	$(t_3 \wedge t_4 \wedge \sim t_5) \vee t_3$
Q_5	$(t_5 \vee t_4) \wedge t_3$	$(t_1 \wedge t_3) \vee (t_4 \wedge t_5)$
Q_6	$t_1 \wedge t_2$	$(t_2 \wedge t_4) \vee t_5$
Q_7	$\sim t_3 \vee (t_4 \wedge t_5)$	$(t_1 \wedge t_3) \vee t_5$
Q_8	$t_1 \wedge t_3$	$(\sim t_1 \wedge t_2) \vee t_3$
Q_9	$t_3 \wedge t_4 \wedge t_5$	$(t_3 \wedge \sim t_4) \vee t_1$
Q_{10}	$(t_1 \vee t_4) \wedge t_3$	$(\sim t_2 \wedge t_3) \vee t_4$
Q_{11}	$(\sim t_4 \wedge t_5)$	$\sim t_1 \wedge t_5$
Q_{12}	$(t_1 \vee t_2 \vee t_3) \wedge t_4$	

Table 3: Case study I - Similarity values S_3 (q_i, q_j), i, j = 1,2,...12, between the queries, as obtained for the document collection (57,775 records) of the Compendex Information System for the period January - September, 1997.

Queries	2	3	4	5	6	7	8	9	10	11	12
1	0.143	1.0	0.034	0.143	1.0	0.031	0.143	0.011	0.143	0.011	0.974
2		0.060	0.321	0.525	0.060	0.252	0.773	0.252	1.0	0.005	0.457
3			0.034	0.143	1.0	0.031	0.143	0.011	0.143	0.011	0.974
4				0.923	0.005	0.356	0.030	0.288	0.309	0.002	0.356
5					0.007	0.106	0.034	0.106	0.775	0.001	0.767
6						0.031	0.143	0.011	0.143	0.011	0.974
7							0.016	0.567	0.567	0.003	0.707
8								0.033	1.0	0.006	0.297
9									1.0	0.005	1.0
10										0.001	0.924
11											1.0

Figures in brackets indicates the actual number of hit documents.

Table 4: Case study II - Similarity values S_3 (q_i, q_j), i, j = 1,2,...11, between the queries, as obtained for the document collection (17,344 records) of the Compendex Information System for the period January - September, 1994.

Queries	2	3	4	5	6	7	8	9	10	11

1	0.755	0.093	0.755	0.739	0.115	0.739	0.755	0.919	0.460	0.009
2		0.035	1.0	0.172	0.040	0.192	1.0	0.172	0.050	0.002
3			0.48	0.18	1.0	0.88	0.48	0.32	0.21	0.3
4				0.097	0.044	0.133	1.0	0.097	0.009	0
5					0.084	1.0	1.0	1.0	0.084	0.004
6						0.997	0.004	0.042	0.006	0.001
7							0.201	(0.011	0.012	0
8								0.097	0.009	0
9									0.030	0.001
10										0.020

Table 5: The Kendall's coefficient of Concordance - W, Chi square - χ^2 and Spearman Rank order Correlation Coefficients for $r_s(S,S^*)$ and $r_s(S, S^{\ominus})$ for all the queries.

	Case Study I				Case Study II			
Query	W	χ^2	$r_s(S, S^*)$	$r_s(S, S^{\ominus})$	W	χ^2	$r_s(S, S^*)$	$r_s(S, S^{\ominus})$
(q_1)	0.72	21.6	0.748	0.488	0.521	14.067	-0.071	0.289
(q_2)	0.69	18.57	0.406	0.556	0.745	20.115	0.909	0.492
(q_3)	0.642	19.26	0.272	0.786	0.689	18.603	0.574	0.111
(q_4)	0.542	16.26	0.271	0.475	0.942	25.434	0.872	0.900
(q_5)	0.803	24.09	0.619	0.890	0.767	20.709	0.480	0.831
(q_6)	0.834	25.02	0.737	0.855	0.570	15.39	0.152	0.244
(q_7)	0.46	13.8	0.065	0.678	0.732	19.764	0.529	0.543
(q_8)	0.833	24.99	0.796	0.865	0.846	22.842	0.726	0.914
(q_9)	0.832	24.96	0.913	0.754	0.624	16.848	0.379	0.787
(q_{10})	0.844	25.32	0.753	0.816	0.575	15.525	0.458	0.486
(q_{11})	0.355	10.65	0	0	0.612	16.524	0	0
(q_{12})	0.725	21.75	0.843	0.538	----	----	----	----

References

1. Czeslaw Danilowicz, 1994. Modeling of User Preferences and Needs in Boolean Retrieval systems. *Information Processing & Management,* 30, 3: 363 - 378.

2. David Goldberg, David Nichols, Brain M. Oki, and Douglas Terry, 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM,* 35, 12: 61 - 70.

3. Degi Young and Ben Shneiderman, 1993. A Graphical Filter/Flow Representation of Boolean Queries: A Prototype Implementation and Evaluation. *Journal of the American Society for Information Science,* 44, 6: 327 - 339.

4. Ellen Voorhees M., 1986. Implementing Agglomerative Hierarchic Clustering Algorithms for use in Document Retrieval. *Information Processing & Management,* 22, 6: 465 - 476.

5. Emtage A. and P. Deutsch, 1992. Archie: An Electronic Directory Service for the Internet. *Proceedings of the Winter USENIX Conference,* pp. 93-110.

6. Gerard Salton, 1968. *Automatic Information Organization and Retrieval.* Mc Graw-Hill Book, Co., New York.

6. Gerard Salton and Chris Buckley, 1990. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science,* 41, 4: 288 - 297.

7. Gerard Salton, Edward A. Fox, and Ellen M. Voorhees, 1985. Advanced Feedback Methods in Informaiton Retrieval. *Journal of the American Society for Information Science,* 36, 3: 200-210.

8. Gerard Salton, Edward A. Fox, Harry Wu, 1983. Extended Boolean Information Retrieval. *Communications of the ACM,* 26, 12: 1022-1036.

9. Jack Minker, Gerald A. Wilson and Barbara H. Zimmerman, 1972. An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System. *Information Storage and Retrieval,* 8: 329-348.

10. Jardine N. and C. J. Van Rijsbergen, 1971. The Use of Hierarchic Clustering in Information Retrieval. *Information Storage Retrieval,* 7: 217-240.

11. Katia Obraczka, Peter B. Danzig and Shih-Hao Li, 1993. Internet Resource Discovery Services. *IEEE Computer Magazine,* 26, 9:8-22.

12. Li Shih-Hao and Peter B. Danzig, 1997. Boolean Similarity Measures for Resource Discovery. *IEEE Transactions on Knowledge and Data Engineering,* 9, 6: 863 - 876.

13. Mark Sheldon A., Andrzej Duda, Ron Weiss, David K. Gifford, 1995. Discover: a Resource Discovery System Based on Content Routing. *Computer Networks and ISDN Systems,* 27: 953-972.

14. Michard, A. , 1982. A New Database Query Language for Non-professional Users: Design Principles and Ergonomic Evaluation. *Behaviour & Information Technology,* 13: 279 - 288.

15. Peter B. Danzig, Shih-Hao Li and Katia Obraczka, 1992. Distrtibuted Indexing of Autonomous Internet Services. *Technical Report, USC TR-92-519,* University of Southern California.

16. Sidney Siegel, 1956. *Non Parametric Statistics for the Behavioral Sciences.* Mcgraw Hill Book Co., New York.

17. Tadeusz Radecki, 1982. Similarity Measures for Boolean Search Request Formulations. *Journal of the American Society for Information Science,* 33, 1: 8 - 17.

18. Tadeusz Radecki, 1985. A theoretical Framework for Defining Similarity Measures for Boolean Search Request Formulations, including some Experimental Results. *Information Processing & Management,* 21, 6: 501-524.

19. Valery I. Frants, Jacob Shapiro, 1991. Algorithm for Automatic construction of Query Formulations in Boolean Form. *Journal of the American Society for Information Science,* 42, 1: 16-26.

20. Van Rijsbergen C. J., 1975. *Information Retrieval.* Butterworth & Co (Publishers) Ltd.

21. Willett, P., 1984. A Note on the Use of Nearest Neighbors for Implementing Single Linkage Document Classifications. *Journal of the American Societ for Information Science,* 35: 149 - 152.