Reasoning with Conceptual Distance
in an Information Retrieval Model

Business Network Section, KTA Research Center, Chung-Gu Chungmoo-Ro 1-Ga 21, Seoul 100-011, Korea and Jin H. Kim
Artifical Intelligence Laboratory
Department of Computer Science, KAIST
P.O. Box 150, Cheongryang, Seoul 130—650, Korea

This paper discusses a reasoning model of information retrieval with a hierarchical thesaurus. The model computes the conceptual distance between a query and an object, both are indexed with weighted terms from a hierarchical thesaurus. The proposed model allows Boolean operators for user queries and edge weights for a hierarchical thesaurus. Experimental results have shown that the proposed model simulates, with surprising accuracy, people in the assessment of conceptual closeness between queries and objects.

1. Introduction

The characteristics of a knowledge based information retrieval (KBIR) system are determined by the knowledge and its representation method. There are various kinds of knowledge which can be extracted from an information retrieval (IR) environment. Among these, a thesaurus, as a type of information structure, is a valuable tool for improving retrieval
effectiveness. A thesaurus provides information on the relationships between the index terms. Among them, hierarchical relationships reveal broader-narrower term relationships from which the conceptual similarity of two given index terms can be deduced. Since IR problem is the identification of those items that contain information pertaining to a user query, the relationships specified in a thesaurus are very valuable in deciding the relevance of an object to a given query. Direct use of this kind of knowledge in deciding the relevance measure will greatly improve the performance of IR system and give a chance to overcome the limitation of the conventional IR model.

2. A Model of Knowledge Based Information Retrieval

Formally, an IR system \( I = \langle T, Q, O, M \rangle \) is defined by the quadruple, a set \( T \) of index terms, a set \( Q \) of queries, a set \( O \) of objects representing the information items, and a matching function \( M \) which computes the conceptual distance between an objects and a query [Kim & Kim 1990].

2.1 HCG as a Knowledge Base

Finite set, \( T \), of \( n \) index terms collectively represents meaningful concepts in the domain under consideration. The set of index terms is prespecified, that is, the controlled vocabulary is used to represent objects as well as queries.

HCG is a weighted hierarchical thesaurus which may be viewed as a directed acyclic graph (DAG) with single source. In HCG, the nodes constitute a set \( T \) of controlled index terms, and edges represent the
"generalization" relation $G$ defined on $T$ with associated weights which reflect the degree of $G$. The notation $t_i \rightarrow G t_j$ indicates that the index term $t_i$ is more general than the index term $t_j$, or alternatively, the index term $t_j$ is more specific than the index term $t_i$. $G$ is irreflexive, asymmetric, and transitive. The hierarchy is based on the partial order imposed by $G$. The edge between two adjacent nodes $t_i$ and $t_j$ has an edge weight $w_{ij}$ reflecting the degree of "generalization" relationship between them. The larger $w_{ij}$ becomes, the less the relatedness between two nodes. According to the research on a comparative study of ranking by human and a computer program, the "is-a" relationship is most often used by humans when evaluating the relationship of objects and queries, and the knowledge embodied in a hierarchical thesaurus is often sufficient to draw inferences for object retrieval [Rada et al. 1985, Rada et al. 1989].

With a variant spreading activation process algorithm [Collins & Loftus 1975], the conceptual distance between any two nodes can be computed by finding shortest path connecting the two nodes and summing the weights of edges along the path between the two. When both nodes are associated with term weight of value 1, the conceptual distance, $d(t_i, t_j)$, between the nodes $t_i$ and $t_j$ is defined as follows:

$$d(t_i, t_j) = d_{ij} = w_{i, s_1} + w_{s_1, s_2} + w_{s_2, s_3} + \cdots + w_{s_n, j},$$  

(1)

where $t_i, t_{s_1}, t_{s_2}, \cdots, t_{s_n}, t_j$ is a sequence of nodes along the shortest path between $t_i$ and $t_j$ in HCG. We assume that the less the conceptual distance between two nodes becomes, the more similar they are. Its validity will be shown through experiments.
2.2 Representation of Query

Each query \( q \) is represented as a legitimate Boolean expression. An weighted index term or its negation is called a literal. A series of literals combined by \textit{and} is called a phrase. Any valid Boolean query can be converted into a disjunctive normal form (DNF), a series of phrases connected together by disjunction. Each phrase, \( ph(q) \), is a conjunctive compound concept and is represented as an \( n \)-dimensional vector, where \( n \) is the cardinality of the set \( T \): \( ph(q) = \langle L_1, L_2, \ldots, L_i, \ldots, L_n \rangle \), where \( L_i \) is \((t_i, w_i)\) or \((\text{not} t_i, w_i)\). Each index term is associated with the weight to represent the importance of the term \((0 \leq w_i \leq 1)\).

2.3 Representation of Object

Each object is represented by an object vector which consists of pairs of index term and weight: \( o = \langle (t_1, w_1), (t_2, w_2), \ldots, (t_i, w_i), \ldots, (t_n, w_n) \rangle \), where \( n \) is the cardinality of the set \( T \). The weight associated with each index term implies the importance of the term for the representation of the object and indicates the degree of fitness of index term to the objects \((0 \leq w_i \leq 1)\).

3. Matching Function Based on Conceptual Distance

We call the matching function \( M \) of our model \textsc{distance}. \textsc{distance} is defined as: \textsc{distance}: \( Q \times O \rightarrow [0, \infty) \). \textsc{distance} assigns a number in the interval \([0, \infty)\) to each pair of query and object. This number is a measure of the conceptual distance.
Reasoning with Conceptual Distance in an Information Retrieval Model 197

Boolean operator or is interpreted as a minimum in computing conceptual distance by the disjunctive minimum rule. \( DISTANCE(q,o) \) is defined as:

\[
DISTANCE(q,o) = \min_{i=1}^{p} DISTANCE(ph_i(q), o)
\]

3.1 Conceptual Distance Between Index Terms

The function \( dw_{ij} \), which computes the conceptual distance between term \( t_i \) associated with weight \( w_i \) and term \( t_j \) associated with weight \( w_j \) is defined as:

\[
dw_{ij} = |w_i - w_j| \cdot \min_{t_i \in A_i} w_{i,b}, \text{ if } t_i = t_j \]

\[
dw_{ij} = d_{ij} + (1 - w_jw_i)\Delta d, \text{ if } t_i \neq t_j,
\]

where \( A_i \) is a set of nodes which are adjacent to \( t_i \), and \( \Delta d \) is a system parameter of the IR system.

3.2 Conceptual Distance on Negated Index Term

Function \( d_{ij} \) represents the conceptual distance between term \( t_i \) with weight of 1 and term \( t_j \) with weight of 1. In order to compute the conceptual distance between \( t_i \) and not \( t_j \), Rada et al. substituted \( T_j \), the set of the farthest nodes from \( t_j \) within the whole semantic net, for not \( t_j \) [Rada et al. 1989]:

\[
\hat{d}_{rij} = \frac{1}{m} \sum_{t \in T_i} d(t_i, t), \text{ where } m \text{ is the cardinality of } T_j
\]

\[
T_j = \{t' \in T | d(t_j, t') = \max_{t \in T} d(t_j, t)\}
\]
However, we substitute the set $t_j^{-1}$ for not $t_j$ by considering the context of a query: $\hat{d}_{ij} = d(t_i, t_j^{-1})$. We call $t_j^{-1}$ the substitution set of not $t_j$. To find $t_j^{-1}$, the negation context subgraph (NCS), in which the context for the negated term is retained, is determined at first [Kim & Kim 1990]. Substitution set of not $t_j, t_j^{-1}$, is defined as:

$$t_j^{-1} = \{ t' \in \text{NCS} | d(t_j, t') = \max_{t \in \text{NCS}} d(t, t_j) \}$$  \hspace{1cm} (6)

$$\hat{d}_{ij} = \min_{t \in t_j^{-1}} (d_{ij})$$  \hspace{1cm} (7)

To preserve the meaning of negation, the conceptual distance of objects indexed by descendants of negated term must be large enough to make those objects judged to be irrelevant. In order to do so, the cut edges connecting the subgraph consisting of negated term and all its successors to the other part of HCG have large edge weight. We call these cut edges "separation edges". In the proposed model, the separation edge weight, $w_{se}$, is set as the length of the longest path within the NCS.

Function $d_{w_{ij}}$ represents the conceptual distance between term $t_i$ with weight of $w_i$ and term not $t_j$ with weight of $w_j$,

$$\hat{d}_{w_{ij}} = |w_i - w_j| \cdot \min_{t \in t_i^{-1}} w_{i,t}, \text{ if } t_i = t_k$$  \hspace{1cm} (8)

$$= \hat{d}_{ij} + (1 - w_i w_j) \Delta d, \text{ if } t_i \neq t_k,$$

where $t_k$ is a term which is an element of $t_j^{-1}$ having the smallest distance from $t_i$.

3.3 Conceptual Distance Between Sets of Index Terms

$DISTANCE(ph(q), o)$ is defined as:

$$DISTANCE(ph(q), o) = \frac{DIS(ph(q), o) + DIS(o, ph(q))}{2}$$  \hspace{1cm} (9)
Reasoning with Conceptual Distance in an Information Retrieval Model

\[
DIS(ph(q), o) = \frac{1}{l} \sum_{t_i \in ph(q)} \min_D W_{ij}
\]

\[
DIS(o, ph(q)) = \frac{1}{m} \sum_{t_i \in o} \min_D W_{ij}
\]

\[D W_{ij} = \hat{d} w_{ij}, \text{ if } t_j \text{ is not negated}\]

\[= d w_{ij}, \text{ if } t_j \text{ is negated},\]

where \(l\) and \(m\) are cardinalities of index terms with positive weight value over zero used for the representation of \(ph(q)\) and \(o\) respectively.

Rada et al. have developed a similar matching function, R-DISTANCE, based on conceptual distance [Rada et al. 1989]. The function covers only the binary indexing scheme. They have defined R-DISTANCE as:

\[
R-DISTANCE(ph(q), o) = \frac{1}{lm} \sum_{t_i \in ph(q)} \sum_{t_j \in o} D_{ij}
\]

\[= 0, \text{ if } ph(q) = o\]

\[D_{ij} = d_{ij}, \text{ if } t_j \text{ is not negated}\]

\[= \hat{d}_{rij}, \text{ if } t_j \text{ is negated}\]

### 4. Evaluation and Discussion

The general method of the evaluation is to compare the performance of DISTANCE to that of human. Since the objective of the knowledge based model is to simulate the performance of human expert, the best reference is people's ranking. Spearman correlation coefficient \(\rho\) [Kendall 1975, p. 8] is used for the evaluation.

#### 4.1 Experiment on DISTANCE

Experiments have been conducted to evaluate R-DISTANCE with
Table 1. Spearman correlation coefficients for CRCS data

<table>
<thead>
<tr>
<th>matching function</th>
<th>correlation for $q_1$</th>
<th>correlation for $q_2$</th>
<th>correlation for $q_3$</th>
<th>correlation for $q_4$</th>
<th>average correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-DISTANCE</td>
<td>0.879</td>
<td>0.740</td>
<td>0.840</td>
<td>0.800</td>
<td>0.815</td>
</tr>
<tr>
<td>DISTANCE</td>
<td>0.867</td>
<td>0.833</td>
<td>0.903</td>
<td>0.817</td>
<td>0.855</td>
</tr>
</tbody>
</table>

several test data [Rada et al. 1985, Rada et al. 1989, McMath et al. 1988]. Test data used in those experiments were good enough to compare the performance of DISTANCE with that of R-DISTANCE for the case of binary indexing. The experimental result with test data from [McMath et al. 1988] is shown in Table 1.

From the result, the correlation of DISTANCE is larger than that of R-DISTANCE on the average. Since R-DISTANCE was shown good to simulate the human ranking, DISTANCE can be claimed better than R-DISTANCE in simulating the human performance.

4.2 Experiment on $d_{ij}$

As explained earlier, DISTANCE adopts a new function for the not operator. To compare the performance of $\hat{d}_{ij}$ in (7) with that of $\hat{d}_{rij}$ in (4), four versions of matching functions were used. These are R-DISTANCE with $\hat{d}_{rij} (m_{11})$, R-DISTANCE with $\hat{d}_{ij} (m_{12})$, DISTANCE with $\hat{d}_{rij}(m_{21})$, and DISTANCE with $\hat{d}_{ij}(m_{22})$.

Test data consist of 5 Boolean queries and 6 articles indexed with CRCS [Sammet & Ralston 1982] terms describing parts of Artificial Intelligence. A Spearman correlation coefficients between each of the four matching functions and the average student's ranking for each query are shown in Table 2. From the result, the average correlation of $m_{12}$ is much larger than that of $m_{21}$. This fact allows us to interpret that $\hat{d}_{ij}$ simulates the human performance more closely than $\hat{d}_{rij}$. In addition,
since the average correlation of $m_{21}$ is larger than that of $m_{11}$ and the average correlation of $m_{22}$ is larger than that of $m_{12}$, we may claim that DISTANCE simulates the human performance more closely than R-DISTANCE.

4.3 Resolution of Discontinuity Near Zero and Counter-intuitive Results

Rada et al. discussed the discontinuity problem in R-DISTANCE through an examplar graph [Rada et al. 1989]. The basic problem seems to be originated from the fact that R-DISTANCE treats all pairs of nodes equally and forces zero property with imposed definition. On the other hand, in DISTANCE function, the zero property is satisfied by nature. More promising is that if we apply DISTANCE to the same example, DISTANCE prevents the counter-intuitive results by resolving the discontinuity problem occurred in R-DISTANCE.

In reality, there exist a mapping of comparable concepts when people usually compare two compound concepts which are represented as conjunctions of the elementary concepts. Utilizing concept mapping in matching function of IR model is very difficult because the 'meaning' has to be understood. For this reason, the closest term (minimum) was taken
as an elementary concept mapping in our DISTANCE function. Taking the closest term seems to be a good heuristic. It solves the discontinuity problem and prevents counter-intuitive results in satisfying the zero property.

In R-DISTANCE, owing to the consideration of all pairwise combinations, the precisely indexed object (the object indexed with more terms) is prone to have larger distance value than the broadly indexed object (the indexed with less terms). However, DISTANCE prevents it. The improvement is due to the fact that DISTANCE only takes its nearest neighbor while R-DISTANCE considers all pairwise combinations.

5. Conclusion

In the proposed IR model, domain-specific relations between index terms are incorporated into the conceptual distance between two objects. HCG plays the role of knowledge base while the matching function DISTANCE plays the role of inference engine. The proposed model retains the advantage of both traditional vector space model and Boolean model allowing Boolean operators in user query and term weight for the representation of the importance of index terms. The experimental results show that the correlation of DISTANCE with people is larger than that of R-DISTANCE on the average and \( \hat{d}_{ij} \) simulates the human performance more closely than \( \hat{d}_{rij} \). However, further experiments are still needed to comprehend the behavior of DISTANCE in handling term weights.
References


Abstract

김영환
한국전신통신공사
연구개발
기술통신망연구실

김진형
한국과학기술원
전산학과
인공지능연구실

본 논문은 계층적 시스테스를 이용한 정보검색 추론모델을 제안하였다. 제안된 모델은 계층적 시스테스를 구성하는 캐인어들과 이들의 가중치로써 표현되는 사용자의 질의어와 정보요소 간의 개념적 거리를 계산한다. 사용자 질의어에 부응하는 연산자를 사용할 수 있도록 하여 검색 요구의 표현력을 향상시켰고, 계층적 시스테스에 에지(edge)가중치를 허용하여 캐인어들간의 상관관계를 보다 정확하게 표현할 수 있도록 하였다. 제안한 모델의 성능 평가를 위한 실험 결과, 인간의 적합도 판정과 상당히 유사함을 알 수 있었다.