UNIT-III

Semantic Networks

Semantic networks are based on the idea that knowledge can be represented by concepts which are linked together by various relationships.

→ A semantic network is simply a set of nodes and arcs.

→ The arcs are labelled for the type of relationship they represent.

→ Factual information about a given node, such as its individual characteristics (color, size, etc.) are often stored in a datastructure called a frame.

→ The individual entries in a frame are called slots [Minsky, 1975].

A frame for a rose can take the form:

```
(rose
      (has-color red)
      (height 2 feet)
      (is-a flower)
)
```

Here the frame 'rose' is a single node in a semantic network containing an is-a link to the node flower. The slots 'has-color' and 'height' store individual properties of the rose.

→ There is a close relationship between a thesaurus and a semantic network. From the standpoint of an information retrieval system, a thesaurus attempts to solve the same mismatch problem by expanding a user query with related terms and hoping that the related terms will match the document. A semantic network subsumes a thesaurus by incorporating links that indicate "is-a-synonym-of" or "is-related-to," but a semantic network can represent more complex information such as an is-a hierarchy which is not found in a thesaurus.

Distance Measures:

To compute the distance between a single node in a semantic network and another node, a spreading activation algorithm is used. A pointer starts at each of the two original nodes and links are followed until an intersection occurs between the two points. The shortest path between the two nodes is used to compute the distance. Note that the simple shortest path algorithm does not apply here because there may be several links that exist between the same two nodes. The distance between nodes a and b is

Distance (a, b) = minimum number of edges separating a and b.

R - distance

$$C_1 = \frac{d(a,t_1) + d(a,t_2) + d(b,t_1) + d(b,t_2) + d(c,t_1) + d(c,t_2)}{6}$$

$$C_2 = \frac{d(e,t_1) + d(e,t_2) + d(f,t_1) + d(f,t_2)}{4}$$

$SC(Q,D)$ is computed now as the MIN (C_1, C_2).

Formally, the R - distance of a disjunctive normal form query Q, and a document D with terms $(t_1, t_2, \ldots t_n)$ and C_{ij}, indicates the j^{th} term in concept i is defined as

$$SC(Q,D) = \min\left(SC_1(C_1,D), SC_1(C_2,D) \ldots SC_1(C_m,D)\right)$$

$$SC_1(C_i,D) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} d(t_i, C_{ij})$$

$SC(Q,D) = 0$ if $Q = D$

K-distance

A subsequent distance measure referred to as the K-distance was developed in [Kim and Kim, 1990].

More formally the distance between terms t_i and t_j is obtained by:

$$d_{ij} = w_{t_i,x_1} + w_{x_1,x_2} + \cdots + w_{x_n,t_j}$$

where the shortest path from t_i to t_j is: $t_i, x_1, x_2, \cdots, t_j$.

To obtain the distance between two sets, A and B, of nodes with weighted arcs, the K-distance measure computes the minimum of the distances are then averaged.

$$C_1 = \frac{\min(d(a,t_1), d(a,t_2)) + \min(d(b,t_1), d(b,t_2)) + \min(d(c,t_1), d(c,t_2))}{3}$$

$$C_2 = \frac{\min(d(e,t_1), d(e,t_2)) + \min(d(f,t_1), d(f,t_2))}{2}$$

$SC(Q,D)$ is still the $MIN(C_1, C_2)$.

The k-distance of a disjunctive normal form query Q and a document D with terms $(t_1, t_2, --, t_n)$ is defined as:

$$SC(Q,D) = \frac{SC_1(Q,D) + SC_1(D,Q)}{2}$$

$$SC_1(Q,D) = \min(SC_2(C_1,D), SC_2(C_2,D), ----, SC_2(C_m,D))$$

$$SC_2(C_i,D) = \frac{1}{n}\left(\sum_{j=1}^{n} \min\left(d(C_{ij}, t_j)\right)\right)$$

$$SC(Q,D) = 0, \text{ if } Q = D.$$

Parsing

The ability to identify a set of tokens to represent a body of text is an essential feature of every information retrieval system. Simply using every token encountered leaves a system vulnerable to fundamental semantic mismatches between a query and a document.

Single Terms

The simplest approach to search documents is to require manual intervention and to assign names of terms to each document. The problem is that it is not always easy to assign keywords that distinctly represent a document. Also, when categorizations are employed - such as the Library of Congress subject headings - it is difficult to stay current within a domain.

Simple phrases

Many TREC systems identify phrases as any pair of terms that are not separated by a stop term, punctuation mark, or special character. Subsequently, infrequently occurring phrases are not stored. In many TREC systems, phrases occurring fewer than 25 times are removed. This dramatically reduces the number of phrases which decreases memory requirements.

Complex phrases

The quest to employ NLP to answer a user query was undertaken since the early 1960's. In fact, NLP systems were often seen as diametrically opposed to information retrieval systems because the NLP systems were trying to understand a document by building a canonical structure that represents the document.

CROSS-LANGUAGE INFORMATION RETRIEVAL

Introduction :-

The Key difference between CLIR and monolingual information retrieval is that the Query and the documents cannot be matched directly. In addition to the Inherent difficulty in matching the inherent style, tone, word usage and other features of the Query with that of the dowment, we must now cross the Language barrier between the Query and the dowmont.

Resources :-

Numerous resources are needed to implement cross-langu retrieval systems. Most approache use bilingual term lists, term dictionaries, a comparable corpus or a parallel corpus.

we also note that even within a single language such as Arabic, there are many different character sets. language processing resources exist to not only detect a language, but also to detect a character set. Cross-language systems often struggle with intricacies involved in working with different character sets within a single language. Unicode was developed to map the character representation for numerous scripts into a single character set, but not all electronic documents are currently stored in Unicode.

*Evaluation:

Different measures are used to evaluate the performance of cross language information retrieval systems. The most obvious is simply to compute the average precision of the cross-language query.

Straightforward techniques typically result in 50% of monolingual performance but the CLIR literature contains results that exceed 100% because of inherent query.

Crossing the Language Barrier

→ what should be translated.? Either the Queries may be translated the documents, or both Queries and documents may be translated, the documents or both Queries and documents may be translated to some internal representation.

→ How should we use a translation.? In other words a single term in language L may map to several terms. Some of or all of these terms.

→ How can we remove spurious translations? Typically, there are spurious translation that can lead to poor retrieval. Techniques exist to remove these translations pruning translations is described in section.

Query Translation :-

Initial work in Query translation was done in the early 1970's where used specified keywords were used to represent document and a dictionary was used to translate English keywords to German Keywords [salton] Query translation approaches use machine translation language Specific Stemmers. the necessary translation of the user Query in language L to the target Query Language 'L'