A CONCEPTUAL FRAMEWORK FOR AUTOMATIC AND DYNAMIC THESAURUS UPDATING IN INFORMATION RETRIEVAL SYSTEMS

M.F. BRUANDET

Laboratoire IMAG
B.P. 53X, 38041 GRENOBLE Cedex (France)

ABSTRACT

This paper aims at presenting a methodology for automatic thesaurus construction in order to help the search of documents and we want to obtain the development of classes for specific topics (for a given corpus) without a priori semantic information. Information contained in the thesaurus lead to new search formulations via automatic and/or user feedback. This presentation even being theoretical is oriented toward a database implementation.

Preliminary remarks

Different strategies used in Information Retrieval Systems must be developped to increase "recall" and "precision"[8,9]. The classic one is the construction of thesaurus. A thesaurus is usually defined as a set of terms (called descriptors) and a set of relations between these terms.

This study is made for an information retrieval system using an inverted file (bitmap, each keyword points to a set of documents containing this keyword). For formulating a request the user defines a set of keywords and boolean operators on this set (for example MISTRAL, GOLEM-PASSAT, STAIRS systems). When entering a document into the database, a module (e.g. PIAF)[4,5] generates stems from the data (several grammatical variants of the same word are reduced to a canonical form). We call this form an item.

Thesaurus construction in the context of local documents

Our object is to find a method for the construction of non-hierarchical relations and the definition of item clusters from these relations. A point to be underlined is that this methodology could efficiently be used only on homogeneous collections of texts. To this purpose, we only consider a database subset : the local set of all documents returned from a given query. The local clustering method makes use of the common occurrences of items within a certain "neighborhood", this method has been studied by R. ATTAR and A.S. FRAENKEL (in "Local feedback in full-text retrieval")[1].

Let be D_ℓ the local set of documents retrieved from a given query and T_ℓ the set of items contained in D_ℓ. We define a metrical function which is inversely proportional to the distance between items in the same sentence. Each item is defined by its coordinates (DN, SN, IN) where DN is the document number, SN the sentence number and IN the item number within a sentence.

For any item $t \in T_\ell$, let $w_t(i)$ be the coordinate of the ith occurrence of t.

For any couple $(s,t) \in T_\ell \times T_\ell$, we define

$d = |w_t(i) - w_s(j)|$ the distance between the ith occurrence of t and jth occurrence of s.

In fact

$$(1)\quad d = |IN_t(i) - IN_s(j)| \text{ with } \begin{cases} DN_t(i) = DN_s(j) \\ SN_t(i) = SN_s(j) \end{cases}$$

Let be F a function of the distance d :

$$(2)\quad F(w_t(i), w_s(j)) = \begin{cases} 1/d \text{ if } w_t(i), w_s(j) \text{ are in the same sentence with } d \leq 20 \\ 0 \text{ otherwise.} \end{cases}$$

For s and $t \in T_\ell$ we define :

$$(3)\quad b(s,t) = \sum_i \sum_j F(w_t(i), w_s(j))$$

where the summation is over all occurrences i and j of s and t.

Remark : $b(s,t) = b(t,s)$.

In order to normalize the function, we take $\mu_R(s,t) = \frac{b(s,t)}{f(t)}$ where f(t) is the number of occurrences of t for all local documents D_ℓ

$$0 \leq \mu_R(s,t) \leq 1.$$

Through this function, we obtain for an item s a reference vector R_s which is a list of items t related to s, such as $\mu_R(s,t)$ is greater (or equal) than a threshold α. These values form an eigen vector : E_{R_s}.

Taking into account new local information in thesaurus updating

Without excluding for the thesaurus the search of hierarchical relations (specific or generic), we try to build a set or a group of items having a notion of "similarity" or "liaison" between themselves. This thesaurus is built as the answers of the used Information Retrieval System are analysed. It must be structured so that the updating should be dynamic and automatic ; the implementation study has not yet been examined. The main problem of updating is to take into account "liaisons", "proximities" or "similarities" between the already registered items in the thesaurus and the new liaisons found after a new query.

For any query, we obtain a set of items related to s. Let be R_s the previous reference vector (μ_{R_s} its associated function) and R'_s the newly calculated vector ($\mu_{R'_s}$ its associated function).

A new reference vector may be calculated from R_s and R'_s using two functions m(s,t) and M(s,t) :

(4) $$m(s,t) = \frac{\text{Min}(\mu_{R_s}(s,t), \mu_{R'_s}(s,t))}{1 - |\mu_{R_s}(s,t) - \mu_{R'_s}(s,t)|}$$

(5) $$M(s,t) = \frac{\text{Max}(\mu_{R_s}(s,t), \mu_{R'_s}(s,t))}{1 + |\mu_{R_s}(s,t) - \mu_{R'_s}(s,t)|}$$

The function M involves all the items t which are related, or not, to s in R_s and R'_s (see Table I). The function m allows us to consider only the items which are both in R_s and in R'_s (see Table I).

One might consider m and M to be respectively the union and intersection of items t related to s.

Table I using the above functions m and M (formulas (4), (5))

$\mu_{R_s}(s,t)$	$\mu_{R'_s}(s,t)$	$m = \frac{\text{Min}(\mu_{R_s}, \mu_{R'_s})}{1-\lvert\mu_{R_s} - \mu_{R'_s}\rvert}$	$M = \frac{\text{Max}(\mu_{R_s}, \mu_{R'_s})}{1+\lvert\mu_{R_s} - \mu_{R'_s}\rvert}$
0	1	indeterminate	0.5
0	0.2	0	0.166
0	0.8	0	0.44
0.1	0.9	0.5	0.5
0.1	0.8	0.33	0.47
0.1	0.7	0.25	0.43
0.1	0.6	0.2	0.40
0.1	0.5	0.16	0.36
0.1	0.4	0.142	0.30
0.1	0.3	0.125	0.25
0.1	0.2	0.11	0.18
0.1	0.1	0.1	0.1
0.5	0.5	0.5	0.5
0.9	0.2	0.66	0.52
0.9	0.4	0.8	0.6
0.9	0.5	0.83	0.64
0.9	0.6	0.85	0.69
0.9	0.8	0.88	0.81

Functions m and M consider the weakest and the strongest bindings between items. Any association between s and t is meaningful only as regard to the "binding strength", that is to say the value of the association function.

Use of the functions m and M for thesaurus construction and updating

For an item x, only the items related to x in several local contexts must be considered in the thesaurus. Thus, it is necessary to keep records of the initial queries into a pseudo-thesaurus. In this pseudo-thesaurus is registered, for any item x, the set of items related to x in one or more local contexts.

Let be

$$PS_x = \{t / \mu_{PS}(x,t) \geq \alpha\}$$

for x belonging to the set of items T, ($T = \cup T_\ell$).

Concerning an item x of T_ℓ, three reference vectors (and their associated functions) can be yielded : R_x, PS_x and T_x which are the sets of items t related to x respectively considered in the treated local context, in one or more local contexts kept in the pseudo-thesaurus, and in the global context kept in the thesaurus.

These sets can be void, also several cases can be encountered :

1) PS_x and T_x are not void

The updating process is performed in three steps :

Step 1 : Treatment of new data

In order to know, if the newly calculated liaisons in R_x already exist in other local context, we compare R_x and PS_x.

Only the common items of these two reference vectors are considered, and we form a temporary reference vector P_x using the function m (formula (4)).

In P_x only items from R_x which are previously related to x in at least one context are retained. The stronger connections are decreased (see Table I) because we can suppose they are only local.

Step 2 : Thesaurus updating procedure

The thesaurus updating is made in two different ways :

(i) if P_x and T_x contain the same items t, only the eigen vector E_{T_x} (of T_x) is modified using the function m (formula (4)) ;

(ii) if the items t in T_x are different from those occuring in P_x, then a new reference vector T_x is constructed combinating the values of functions μ_{T_x} and μ_{PS_x} using M (formula (5)).

Remarks :
- We do not calculate the new association function between two items for T_x with m (formula (4)), because we do not introduce new items related to x in the thesaurus, when new items appear in several local contexts.
- The function M uses the common or not common items and introduces in the thesaurus the new items, which are related to x in at least two local contexts.

Step 3 : Pseudo-thesaurus updating procedure
The pseudo-thesaurus updating must take into account the new items occuring in R_x. The new association function for PS_x is calculated from the association function μ_{R_x} and the old association function μ_{PS_x} using M (formula (5)).

2) PS_x and T_x are void

This case corresponds to the situation where x is never appeared in any local context. We create the reference vectors PS_x in the pseudo-thesaurus and R_x with the association function μ_{R_x} ($PS_x = R_x$). No information about x is kept in the thesaurus ($T_x = \emptyset$).

3) PS_x is not void and T_x is void

This corresponds to the case where x is already appeared in only one local context. If $R_x \neq \emptyset$, then we can build the initial reference vector T_x in the thesaurus. We use the association function m (formula (4)) calculated from the values of association functions μ_{R_x} and μ_{PS_x} (respectively contained in the eigen vectors E_{R_x} and E_{PS_x}).

The present experimentation exhibits among the items related to x in T_x (initial step) local synonyms, some global synonyms and many parasistic items. After a few thesaurus updatings the values of the association function for parasistic items rapidly decrease, and the values for local and global synonyms increase. It is clear that reliability of such a thesaurus can be reached only after a large number of queries. In such a situation new updating procedures might be considered so that new parasistic items should not be introduced in T_x (thus breaking the stability of T_x).

Global treatment of thesaurus

Let be T the large set of items registered in the thesaurus. In order to classify T (i.e. to split T into classes of similar items), we consider the couple of reference vectors T_x and T_y (so E_{T_x} and E_{T_y}) for any items x and y.

Let be r(x,y) a similarity measure :

$$(6)\quad r(x,y) = \frac{\sum_T \mathrm{Min}(\mu_{T_x}, \mu_{T_y})}{\sum_T \mathrm{Max}(\mu_{T_x}, \mu_{T_y})}$$

(7) d(x,y) = 1-r(x,y) is a pseudo-distance whose range is [0,1].

We can use an association matrix (i.e. term-term matrix) between items and found a partition of T in equivalence classes. Moreover, this method hardly applies to a great many items and does not seem realistic for a large scale dictionnary (6000 or 10000 items, for example) which are common in information retrieval field. To overcome this drawback, we may try to build up the global association matrix from the local ones. Some ideas have been suggested[2] using the fuzzy sets theory[6,13] but there are still theoretical approaches.

Feedback query processing

Number of papers are related to the feedback query processing[1,8,12] and our approach is similar.

We think to adopt the following strategy, though we lack practical results to assert better "score" on queries.

After a query we have therefore a set R_s of items related to s (for each $s \in T_\ell$) and a partition of T_ℓ into equivalence classes Γ_j. In the thesaurus we might have both a set T_s (items related to s) and a partition of the global set T into equivalence classes C_i.

Several strategies can be used, they are detailed in an other paper[4]. We can use only local context, global context or both global and local context. We summarize some of the solutions below :

1) use of only global context
A query is automatically generated with t instead of x when t belongs to the reference vector T_x and $\mu_{T_x}(x,t)$ is greater or equal than a threshold α.

If the user agrees, a new query is generated with t when x and t are equivalent in the thesaurus.

2) use of both local and global context
When an item t is considered as "similar" to x both in local context (R_x) and in global context (T_x) and $\mu_{R_x}(x,t) \leq \mu_{T_x}(x,t)$, t automatically replaces x in the query. When R_x and T_x have common items, we can purpose to the user new queries with item t appearing in T_x but not in R_x ($\mu_{T_x}(x,t) \geq \alpha$).

As previously mentioned we can use the same strategy using the local equivalence classes Γ_j and global equivalence classes C_i (automatic feedback query processing with $x \in C_i \cap \Gamma_j$, and under user control with $x \in C_i$ but $x \notin C_i \cap \Gamma_j$ and $C_i \cap \Gamma_j \neq \emptyset$).

In this last case, we can think global synonymies allow to retrieve new documents originally left out.

From the previous analysis, it seems that the best strategies should be those using both local and global contexts, but this needs to be verified.

Conclusions

We conclude from present experimentation on small number of french texts that the thesaurus updating method shall give horizontal thesaurus relations.

Moreover unexpected relation between items should appear in the thesaurus, that is association which strongly reflects the corpus' content and which could not a priori be established and enhanced.

The methodology presented above does not exclude any further intervention on the thesaurus to refine semantic information about some particular cases, such as modifying values of the association function for some items, enriching definition of synonyms.

Our next goal for such a design of the thesaurus is twofold :

1) we wish to make possible non boolean queries through the use of fuzzy keywords and subsequent improvement of dialogue ;

2) we wish to cluster documents with a dynamic indexing mechanism.

REFERENCES

1 R. ATTAR & A.S. FRAENKEL
Local feedback in full text retrieval systems.
Journal of ACM, vol.20, n°3, pp. 397-417, July 1977.

2 M.F. BRUANDET
A propos de la construction automatique d'un thesaurus flou dans un système de recherche d'information (système documentaire).
Internal research report IMAG Grenoble, Juin 1980.

3 M.F. BRUANDET
A conceptual framework for automatic and dynamic thesaurus updating and for feedback query processing.
Processing of SECOND INTERNATIONAL CONFERENCE ON DATA BASES IN THE HUMANITIES AND SOCIAL SCIENCES, Madrid, Juin 1980.

4 J. COURTIN
Algorithmes pour le traitement interactif des langues naturelles.
Thèse d'Etat soutenue à l'Université Scientifique et Médicale de Grenoble, INPG, Octobre 1977.

5 E. GRANDJEAN
Projet PIAF - Application à la documentation automatique : définition et utilisation du produit prototype PIAFDOC.
Internal research report, IMAG Grenoble, 1979.

6 T. RADECKI
Mathematical model of information retrieval system based on the concept of fuzzy thesaurus.
Information processing and management, vol.12, pp. 313-318, Pergamon Press, 1976.

7 L. REISINGER
On fuzzy thesaurus.
COMPSTAT/4 - Proc. Symp. Computational statistics, Bruckman b, Fershl l, Schmetterer - Vienna Physics Verlag.

8 G. SALTON
The smart retrieval system, experiments in automatic.
Document processing (ch.21 - the use of statistical significance in relevance feedback. J.S. Brown, P.D. Reilly), Prentice Hall, 1971.

9 G. SALTON
Dynamic information processing.
Prentice Hall 1975.

10 G. SALTON and D. BERGMARK
Clustered file generation and its application to computer Science taxonomies.
IFIP Information processing 77, pp. 441-447, North Holland publishing company.

11 W. SILVERT
Symmetric summation : a class of operations on fuzzy sets.
IEEE Trans. SMC, 1979.

12 C.T. YU, M.K. SIU
Effective automatic indexing using term addition and deletion.
Journal of ACM, vol.12, n°2, April 1978, pp. 210-225.

13 L.A. ZADEH
Fuzzy sets, Information and control.
pp. 338-353, 1965.