BUILDING A LEXICAL DOMAIN MAP FROM TEXT CORPORA

Tomek Strzalkowski

Courant Institute of Mathematical Sciences, New York University
715 Broadway, rm. 704, New York, NY 10003, tomek@cs.nyu.edu

SUMMARY

In information retrieval the task is to extract from the database all and only the documents which are relevant to a user query, even when the query and the documents use little common vocabulary. In this paper we discuss the problem of automatic generation of lexical relations between words and phrases from large text corpora and their application to automatic query expansion in information retrieval. Reported here are some preliminary results and observations from the experiments with a 85 million word Wall Street Journal database and a 45 million word San Jose Mercury News database (parts of 0.5 billion word TIPSTER/TREC database).

INTRODUCTION

The task of information retrieval is to extract *relevant* documents from large collection of documents in response to a user's query. When the documents contain primarily unrestricted text (e.g., newspaper articles, legal documents, etc.) the relevance of a document is established through 'full-text' retrieval. This has been usually accomplished by identifying key terms in the documents (the process known as 'indexing') which could then be matched against terms in queries (Salton, 1989). The effectiveness of any such term-based approach is directly related to the accuracy with which a set of terms represents the content of a document, as well as how well it contrasts a given document with respect to other documents. In other words, we are looking for a representation R such that for any text items $D1$ and $D2$, $R(D1) = R(D2)$ iff $meaning(D1) = meaning(D2)$, at an appropriate level of abstraction (which may depend on types and character of anticipated queries).

For all kinds of terms that can be assigned to the representation of a document, e.g., words, operator-argument pairs, fixed phrases, and proper names, various levels of "regularization" are needed to assure that syntactic or lexical variations of input do not obscure underlying semantic uniformity. Without actually doing semantic analysis, this kind of normalization can be achieved through the following processes:[1]

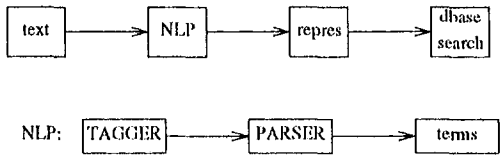
(1) morphological stemming: e.g., *retrieving* is reduced to *retriev*;

(2) lexicon-based word normalization: e.g., *retrieval* is reduced to *retrieve*;

(3) operator-argument representation of phrases: e.g., *information retrieval*, *retrieving of information*, and *retrieve relevant information* are all assigned the same representation, *retrieve+information*;

(4) context-based term clustering into synonymy classes and subsumption hierarchies: e.g., *take-over* is a kind of *acquisition* (in business), and *Fortran* is a *programming language*.

We have established the general architecture of a NLP-IR system that accommodates these considerations. In a general view of this design, depicted schematically below, an advanced NLP module is inserted between the textual input (new documents, user queries) and the database search engine (in our case, NIST's PRISE system).



This design has already shown some promise in producing significantly better performance than the base statistical system (Strzalkowski, 1993). Its practical significance stems in no small part from the use of a fast and robust parser, TTP,[2] which can process unrestricted text at speeds below 0.2 sec per sentence. TTP's output is a regularized representation of each sentence which reflects logical predicate-argument structure, e.g., logical subject and logical objects are identified depending upon the main verb subcategorization frame. For example, the verb *abide* has, among others, a subcategorization frame in which the object is a prepositional phrase with *by*, i.e.,

ABIDE: *subject* NP *object* PREP by NP

Subcategorization information is read from the on-line Oxford Advanced Learner's Dictionary (OALD) which TTP uses.

[1] An alternative, but less efficient method is to generate all variants (lexical, syntactic, etc.) of words/phrases in the queries (Sparck-Jones & Tait, 1984).

[2] TTP stands for Tagged Text Parser, and it has been described in detail in (Strzalkowski, 1992) and evaluated in (Strzalkowski & Scheyen, 1993).

HEAD-MODIFIER STRUCTURES

TTP parse structures are passed to the phrase extraction module where head+modifier (including predicate+argument) pairs are extracted and collected into occurrence patterns. The following types of head+modifier pairs are extracted:

(1) a head noun and its left adjective or noun adjunct,

(2) a head noun and the head of its right adjunct,

(3) the main verb of a clause and the head of its object phrase.

These types of pairs account for most of the syntactic variants for relating two words (or simple phrases) into pairs carrying compatible semantic content. For example, the pair *retrieve+information* will be extracted from any of the following fragments: *information retrieval system; retrieval of information from databases*; and *information that can be retrieved by a user-controlled interactive search process.*[3]

Figure 1 shows TTP parse and head+modifier pairs extracted. Whenever multiple-noun strings (two nouns plus another noun or adjective) are present, they need to be structurally disambiguated before any pairs can be extracted. This is accomplished using statistically-based preferences, e.g., *world+third* is preferred to either *country+world* or *country+third* when extracted from *third world country*. If such preferences cannot be computed, all alternatives are discarded to avoid noisy input to clustering programs.

[San Jose Mercury News 08/30/91 Business Section]
For McCaw, it would have hurt the company's strategy of building a seamless national cellular network.

```
[assert,
  [[will_aux],[[perf,[have]],
  [[verb,[hurt]],
    [subject,[np,[n,it]]],
    [object,[np,[n,strategy],[t_pos,the],
          [n_pos,[poss,[n,company]]],
          [of,
           [[verb,[build]],
             [subject,anyone],
             [object,[np,[n,network],[t_pos,a],
                   [adj,[seamless]],
                   [adj,[national]],
                   [adj,[cellular]]]]]]]]],
  [for,[np,[name,[mccaw]]]]].
```

EXTRACTED PAIRS:

hurt+strategy	strategy+company
build+network	network+cellular
network+national	network+seamless

Figure 1. Extracting Head+Modifier pairs from parsed sentences.

TERM CORRELATIONS FROM TEXT

Head-modifier pairs serve as occurrence contexts for terms included in them: both single words (as shown in Figure 1) and other pairs (in case of nested pairs, e.g., *country+[world+third]*). If two terms tend to be modified with a number of common modifiers but otherwise appear in few distinct contexts, we assign them a similarity coefficient, a real number between 0 and 1. The similarity is determined by comparing distribution characteristics for both terms within the corpus: in general we will credit high-content terms appearing in multiple identical contexts, provided that these contexts are not too commonplace.[4] Figure 2 shows examples of terms sharing a number of common contexts along with frequencies of occurrence in a 250 MByte subset of Wall Street Journal database. A *head context* is when two distinct modifiers are attached to the same head element; a *mod context* is when the same term modifies two distinct heads.

To compute term similarities we used a variant of weighted Jaccard's measure described in e.g., (Grefenstette, 1992):[5]

TERM1	TERM2	COMM CNTXT HEAD	MOD	FRQ1	FRQ2
vice	deputy	president		9295	29
		chairman		1007	146
		director		6	158
		minister		37	17
		premier		7	8
man	boy	story		9	3
		club		6	4
		age		18	3
		mother		4	5
			bad	4	4
			young	258	12
			older	18	4

Figure 2. Example pairs of related terms.

[3] subject+verb pairs are also extracted but these are not used in the lexical clustering procedure described here.

[4] It would not be appropriate to predict similarity between *language* and *logarithm* on the basis of their co-occurrence with *natural*.

[5] In another series of experiments (Strzalkowski & Vauthey, 1992) we used a Mutual Information based classification formula (e.g., Church and Hanks, 1990; Hindle, 1990), but we found it less effective for diverse databases, such as WSJ.

$$SIM(x_1,x_2) = \frac{\sum_{att} MIN(W([x,att]),W([y,att])}{\sum_{att} MAX(W([x,att]),W([y,att])}$$

with

$$W([x,y]) = GEW(x)*log(f_{x,y})$$

$$GEW(x) = 1 + \sum_{y}\left[\frac{\frac{f_{x,y}}{n_y} * \log\left[\frac{f_{x,y}}{n_y}\right]}{log(N)}\right]$$

In the above, $f_{x,y}$ stands for absolute frequency of pair $[x,y]$ in the corpus, n_y is the frequency of term y, and N is the number of single-word terms.

In order to generate better similarities, we require that words x_1 and x_2 appear in at least M *distinct* common contexts, where a common context is a couple of pairs $[x_1,y]$ and $[x_2,y]$, or $[y,x_1]$ and $[y,x_2]$ such that they each occurred at least K times. Thus, *banana* and *Baltic* will not be considered for similarity relation on the basis of their occurrences in the common context of *republic*, no matter how frequent, unless there are $M-1$ other such common contexts comparably frequent (there wasn't any in TREC's WSJ database). For smaller or narrow domain databases M=2 is usually sufficient, e.g., CACM database of computer science abstracts. For large databases covering a diverse subject matter, like WSJ or SJMN (San Jose Mercury News), we used $M \geq 5$.[6] This, however, turned out not to be sufficient. We would still generate fairly strong similarity links between terms such as *aerospace* and *pharmaceutical* where 6 and more common contexts were found, even after a number of common contexts, such as *company* or *market*, have already been rejected because they were paired with too many different words, and thus had a dispersion ratio too high. The remaining common contexts are listed in Figure 3, along with their GEW scores, all occurring at the head (left) position of a pair.

CONTEXT	GEW	frequency with	
		aerospace	*pharmaceutical*
firm	0.58	9	22
industry	0.51	84	56
sector	0.61	5	9
concern	0.50	130	115
analyst	0.62	23	8
division	0.53	36	28
giant	0.62	15	12

Figure 3. Common (head) contexts for *aerospace* and *pharmaceutical*.

[6] For example *banana* and *Dominican* were found to have two common contexts: *republic* and *plant*, although this second occurred in apparently different senses in *Dominican plant* and *banana plant*.

When analyzing Figure 3, we should note that while some of the GEW weights are quite low (GEW takes values between 0 and 1), thus indicating a low importance context, the frequencies with which these contexts occurred with both terms were high and balanced on both sides (e.g., *concern*), thus adding to the strength of association. To filter out such cases we established thresholds for admissible values of GEW factor, and disregarded contexts with entropy weights falling below the threshold. In the most recent experiments with WSJ texts, we found that 0.6 is a good threshold. We also observed that clustering head terms using their modifiers as contexts converges faster and gives generally more reliable links than when mod terms are clustered using heads as context (e.g., in the above example). In our experiment with the WSJ database, we found that an occurrence of a common head context needs to be considered as contributing less to the total context count than an occurrence of a common mod context: we used 0.6 and 1, respectively. Using this formula, terms *man* and *boy* in Figure 2 share 5.4 contexts (4 head contexts and 3 mod contexts).

Initially, term similarities are organized into clusters around a centroid term. Figure 4 shows top 10 elements (sorted by similarity value) of the cluster for *president*. Note that in this case the SIM value drops suddenly after the second element of the cluster. Changes in SIM value are used to determine cut-off points for clusters. The role of GTS factor will be explained later. Sample clusters obtained from approx. 250 MByte (42 million words) subset of WSJ (years 1990-1992) are given in Table 1.

It may be worth pointing out that the similarities are calculated using term co-occurrences in syntactic rather than in document-size contexts, the latter being the usual practice in non-linguistic clustering (e.g., Sparck Jones and Barber, 1971; Crouch, 1988; Lewis and Croft, 1990). Although the two methods of term clustering may be considered mutually complementary in certain situations, we believe that more and stronger associations can be obtained through syntactic-context clustering, given sufficient amount of data and a reasonably accurate syn-

CENTROID	TERM	SIM	GTS
president			0.0011
	director	0.2481	0.0017
	chairman	0.2449	0.0028
	office	0.1689	0.0010
	manage	0.1656	0.0007
	executive	0.1626	0.0012
	official	0.1612	0.0008
	head	0.1564	0.0018
	member	0.1506	0.0014
	lead	0.1311	0.0009

Figure 4. A cluster for *president*.

word	cluster
takeover	*merge, buy-out, acquire, bid*
benefit	*compensate, aid, expense*
capital	*cash, fund, money*
staff	*personnel, employee, force*
attract	*lure, draw, woo*
sensitive	*crucial, difficult, critical*
speculate	*rumor, uncertainty, tension*
president	*director, chairman*
vice	*deputy*
outlook	*forecast, prospect, trend*
law	*rule, policy, legislate, bill*
earnings	*profit, revenue, income*
portfolio	*asset, invest, loan*
inflate	*growth, demand, earnings*
industry	*business, company, market*
growth	*increase, rise, gain*
firm	*bank, concern, group, unit*
environ	*climate, condition, situation*
debt	*loan, secure, bond*
lawyer	*attorney*
counsel	*attorney, administrator, secretary*
compute	*machine, software, equipment*
competitor	*rival, competition, buyer*
alliance	*partnership, venture, consortium*
big	*large, major, huge, significant*
fight	*battle, attack, war, challenge*
base	*facile, source, reserve, support*
shareholder	*creditor, customer, client investor, stockholder*

Table 1. Selected clusters obtained from syntactic contexts, derived from approx. 40 million words of WSJ text, with weighted Jaccard formula.

tactic parser.[7]

[7] Non-syntactic contexts cross sentence boundaries with no fuss, which is helpful with short, succinct documents (such as CACM abstracts), but less so with longer texts; see also (Grishman et al., 1986).

QUERY EXPANSION

Similarity relations are used to expand user queries with new terms, in an attempt to make the final search query more comprehensive (adding synonyms) and/or more pointed (adding specializations). It follows that not all similarity relations will be equally useful in query expansion, for instance, complementary and antonymous relations like the one between *Australian* and *Canadian*, *accept* and *reject*, or even generalizations like from *aerospace* to *industry* may actually harm system's performance, since we may end up retrieving many irrelevant documents. On the other hand, database search is likely to miss relevant documents if we overlook the fact that *vice director* can also be *deputy director*, or that *takeover* can also be *merge*, *buy-out*, or *acquisition*. We noted that an average set of similarities generated from a text corpus contains about as many "good" relations (synonymy, specialization) as "bad" relations (antonymy, complementation, generalization), as seen from the query expansion viewpoint. Therefore any attempt to separate these two classes and to increase the proportion of "good" relations should result in improved retrieval. This has indeed been confirmed in our experiments where a relatively crude filter has visibly increased retrieval precision.

In order to create an appropriate filter, we devised a global term specificity measure (GTS) which is calculated for each term across all contexts in which it occurs. The general philosophy here is that a more specific word/phrase would have a more limited use, i.e., a more specific term would appear in fewer *distinct* contexts. In this respect, GTS is similar to the standard *inverted document frequency (idf)* measure except that term frequency is measured over syntactic units rather than document size units. Terms with higher GTS values are generally considered more specific, but the specificity comparison is only meaningful for terms which are already known to be similar. We believe that measuring term specificity over document-size contexts (e.g., Sparck Jones, 1972) may not be appropriate in this case. In particular, syntax-based contexts allow for processing texts without any internal document structure.

The new function is calculated according to the following formula:

$$GTS(w) = \begin{cases} IC_L(w) * IC_R(w) & \text{if } both\ exist \\ IC_R(w) & \text{if } only\ IC_R(w)\ exists \\ IC_L(w) & otherwise \end{cases}$$

where (with n_w, $d_w > 0$):

$$IC_L(w) = IC([w,\_]) = \frac{n_w}{d_w(n_w+d_w-1)}$$

$$IC_R(w) = IC([\_,w]) = \frac{n_w}{d_w(n_w+d_w-1)}$$

In the above, d_w is *dispersion* of term w understood as the number of distinct contexts in which w is found. For any two terms w_1 and w_2, and a constant $\delta_1 > 1$, if $GTS(w_2) \geq \delta_1 * GTS(w_1)$ then w_2 is considered more specific than w_1. In addition, if $SIM_{norm}(w_1,w_2) = \sigma > 0_1$, where 0_1 is an empirically

established threshold, then w_2 can be added to the query containing term w_1 with weight $\sigma * \omega$,[8] where ω is the weight w_2 would have if it were present in the query. Similarly, if $GTS(w_2) \leq \delta_2 * GTS(w_1)$ and $SIM_{norm}(w_1,w_2) = \sigma > \theta_2$ (with $\delta_2 < \delta_1$ and $\theta_1 < \theta_2$) then we may consider w_2 as synonymous to w_1. All other relations are discarded. For example, the following were obtained from the WSJ training database:

$GTS(takeover) = 0.00145576$
$GTS(merge) = 0.00094518$
$GTS(buy\text{-}out) = 0.00272580$
$GTS(acquire) = 0.00057906$

with

$SIM(takeover,merge) = 0.190444$
$SIM(takeover,buy\text{-}out) = 0.157410$
$SIM(takeover,acquire) = 0.139497$
$SIM(merge,buy\text{-}out) = 0.133800$
$SIM(merge,acquire) = 0.263772$
$SIM(buy\text{-}out,acquire) = 0.109106$

Therefore both *takeover* and *buy-out* can be used to specialize *merge* or *acquire*. With this filter, the relationships between *takeover* and *buy-out* and between *merge* and *acquire* are either both discarded or accepted as synonymous. At this time we are unable to tell synonymous or near synonymous relationships from those which are primarily complementary, e.g., *man* and *woman*.

Filtered similarity relations create a domain map of terms. At present it may contain only two types of links: equivalence (synonymy and near-synonymy) and subsumption (specification). Figure 5 shows a small fragment of such map derived from lexical relation computed from WSJ database. The domain map is used to expand user queries with related terms, either automatically or in a feedback mode by showing the user appropriate parts of the map.

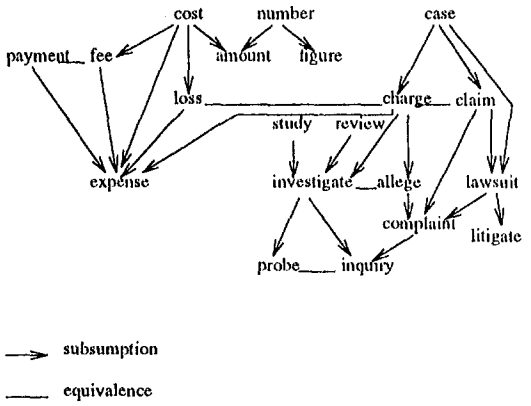


Figure 5. A fragment of the domain map network. Note the emerging senses of 'charge' as 'expense' and 'allege'.

[8] For TREC-2 we used θ=0.2; δ varied between 10 and 100.

We should add that the query expansion (in the sense considered here, though not quite in the same way) has been used in information retrieval research before (e.g., Sparck Jones and Tait, 1984; Harman, 1988), usually with mixed results. The main difference between the current approach and those previous attempts is that we use lexico-semantic evidence for selecting extra terms, while they relied on term co-occurrence within the same documents. In fact we consider these to methods complementary with the latter being more appropriate for automatic relevance feedback. An alternative query expansion to is to use term clusters to create new terms, "metaterms", and use them to index the database instead (e.g., Crouch, 1988; Lewis and Croft, 1990). We found that the query expansion approach gives the system more flexibility, for instance, by making room for hypertext-style topic exploration via user feedback.

CONCLUSIONS

We discussed selected aspects our information retrieval system consisting of an advanced NLP module and a 'standard' statistical core engine. In this paper we concentrated on the problem of automatic generation of lexical correlations among terms which (along with appropriate weighting scheme) represent the content of both the database documents and the user queries. Since a successful retrieval relies on actual term matches between the queries and the documents, it is essential that any lexical alternatives of describing a given topic are taken into account. In our system this is achieved through the expansion of user's queries with related terms: we add equivalent and more specific terms. Lexical relations between terms are calculated directly from the database and stored in the form of a domain map, which thus acts as a domain-specific thesaurus. Query expansion can be done in the user-feedback mode (with user's assistance) or automatically. In this latter case, local context is explored to assure meaningful expansions, i.e., to prevent e.g., expanding 'charge' with 'expense' when 'allege' or 'blame' is meant, as in the following example query:

> Documents will report on corruption, incompetence, or inefficiency in the management of the United Nation's staff. Allegations of management failings, as well as retorts to such *charges* are relevant.

Many problems remain, however, we attempted to demonstrate that the architecture described here is nonetheless viable and has practical significance. More advanced NLP techniques (including semantic analysis) may prove to be still more effective, in the future, however their enormous cost limits any experimental evidence to small scale tests (e.g., Mauldin, 1991).

ACKNOWLEDGEMENTS

We would like to thank Donna Harman of NIST for making her PRISE system available to us. We would also like to thank Ralph Weischedel and Heidi Fox of BBN for providing and assisting in the use of the part of speech tagger. This paper is based upon work supported by the Advanced Research Project Agency under Contract

N00014-90-J-1851 from the Office of Naval Research, under Contract N00600-88-D-3717 from PRC Inc., and the National Science Foundation under Grant IRI-93-02615. We also acknowledge support from the Canadian Institute for Robotics and Intelligent Systems (IRIS).

REFERENCES

Church, Kenneth Ward and Hanks, Patrick. 1990. "Word association norms, mutual information, and lexicography." *Computational Linguistics*, 16(1), MIT Press, pp. 22-29.

Crouch, Carolyn J. 1988. "A cluster-based approach to thesaurus construction." Proceedings of ACM SIGIR-88, pp. 309-320.

Grefenstette, Gregory. 1992. "Use of Syntactic Context To Produce Term Association Lists for Text Retrieval." Proceedings of SIGIR-92, Copenhagen, Denmark. pp. 89-97.

Grishman, Ralph, Lynette Hirschman, and Ngo T. Nhan. 1986. "Discovery procedures for sublanguage selectional patterns: initial experiments". *Computational Linguistics*, 12(3), pp. 205-215.

Harman, Donna. 1988. "Towards interactive query expansion." Proceedings of ACM SIGIR-88, pp. 321-331.

Hindle, Donald. 1990. "Noun classification from predicate-argument structures." Proc. 28 Meeting of the ACL, Pittsburgh, PA, pp. 268-275.

Lewis, David D. and W. Bruce Croft. 1990. "Term Clustering of Syntactic Phrases". Proceedings of ACM SIGIR-90, pp. 385-405.

Mauldin, Michael. 1991. "Retrieval Performance in Ferret: A Conceptual Information Retrieval System." Proceedings of ACM SIGIR-91, pp. 347-355.

Salton, Gerard. 1989. *Automatic Text Processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, MA.

Sparck Jones, Karen. 1972. "Statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation*, 28(1), pp. 11-20.

Sparck Jones, K. and E. O. Barber. 1971. "What makes automatic keyword classification effective?" *Journal of the American Society for Information Science*, May-June, pp. 166-175.

Sparck Jones, K. and J. I. Tait. 1984. "Automatic search term variant generation." *Journal of Documentation*, 40(1), pp. 50-66.

Strzalkowski, Tomek and Barbara Vauthey. 1992. "Information Retrieval Using Robust Natural Language Processing." Proc. of the 30th ACL Meeting, Newark, DE, June-July. pp. 104-111.

Strzalkowski, Tomek. 1992. "TTP: A Fast and Robust Parser for Natural Language." Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, July 1992. pp. 198-204.

Strzalkowski, Tomek. 1993. "Robust Text Processing in Automated Information Retrieval." Proc. of ACL-sponsored workshop on Very Large Corpora. Ohio State Univ. Columbus, June 22.

Strzalkowski, Tomek. 1994. "Document Representation in Natural Language Text Retrieval." To appear in proceedings of ARPA Human Language Technology Workshop, Princeton, NJ. March 8-11.

Strzalkowski, Tomek and Jose Perez-Carballo. 1994. "Recent Developments in Natural Language Text Retrieval." To appear in proceedings of Second Text Retrieval Conference (TREC-2), Gaithersburg, Md, August 30 - September 1, 1993.

Strzalkowski, Tomek, and Peter Scheyen. 1993. "Evaluation of TTP Parser: a preliminary report." Proceedings of International Workshop on Parsing Technologies (IWPT-93), Tilburg, Netherlands and Durbuy, Belgium, August 10-13.

APPENDIX: An example query

The following is an example information request (based on TREC's topic 113) and the resulting query. Except for its inverted document frequency score, each term has a "confidence level" weight which is set to 1.0 if the term is found in the user's query, and is less than 1.0 if the term is added through an expansion from the domain map. Only non-negated terms with *idf* of 6.0 or greater are included.

<title> New Space Satellite Applications

<desc> Document will report on non-traditional applications of space satellite technology.

<narr> A relevant document will discuss more recent or emerging applications of space satellite technology. NOT relevant are such "traditional" or early satellite age usages as INTELSAT transmission of voice and data communications for telephone companies or program feeds for established television networks. Also NOT relevant are such established uses of satellites as military communications, earth mineral resource mapping, and support of weather forecasting. A few examples of newer applications are the building of private satellite networks for transfer of business data, facsimile transmission of newspapers to be printed in multiple locations, and direct broadcasting of TV signals. The underlying purpose of this topic is to collect information on recent or emerging trends in the application of space satellite technology.

TERM	*IDF*	*WEIGHT*
apply+equip	18.402237	0.458666
satellite+latest	18.402237	0.254058
television+signal	18.402237	0.359777
television+direct	18.402237	0.359777
apply+equip	18.402237	0.458666
broadcast+direct	16.402237	1.000000
location+multiple	16.402237	1.000000
broadcast+signal	16.080309	1.000000
support+forecast	15.817275	1.000000
data+business	15.817275	1.000000
forecast+internal	15.402238	0.283029
transfer+inform	15.232312	0.511940
transfer+data	14.817275	1.000000
figure+business	14.594883	0.453631

technology+satellite	14.495347	1.000000
transmit+facsimile	14.402238	1.000000
equip+satellite	14.232312	0.458666
signal+broadcast	13.701797	0.441993
signal+tv	13.701797	1.000000
signal+television	13.594883	0.813987
news+business	13.495347	0.352291
network+satellite	13.154310	1.000000
develop+network	12.942806	0.409144
non+traditional	12.758382	1.000000
inform+business	12.729813	0.511940
apply+technology	12.471500	1.000000
build+network	11.212413	1.000000
facsimile	10.217362	1.000000
usage	9.902391	1.000000
newer	9.306841	1.000000
elderly	8.202565	0.361246
feed	7.802325	1.000000
satellite	7.567767	1.000000
underly	7.370192	1.000000
transmit	7.299606	1.000000
multiple	7.241736	1.000000
broadcast	7.019614	1.000000
location	6.992316	1.000000
print	6.351709	1.000000
space	6.226376	1.000000
transfer	6.155497	1.000000
collect	6.126113	1.000000
signal	6.080873	1.000000
phone	6.072441	0.663414
tv	6.003761	1.000000