CONCEPT & EFFECTIVENES of the COVER-COEFFICIENT-BASED CLUSTERING METHODOLOGY (C^3M) FOR TEXT DATABASES

* IRSs PERFORM RETRIEVALS ON DOCUMENT REPRESENTATIVES.

$$D = \begin{bmatrix} x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix}$$

→ SINGLE DOCUMENT

↓ TERM

* $d_{ij} (1 \leq i \leq m, 1 \leq j \leq n)$ indicates, the importance of term$j$ (t_j) in document i (d_i).
* Indexing can be binary or weighted.

* QUERY PROCESSING RETRIEVES DOCUMENTS THAT ARE RELEVANT TO THE USER REQUEST.
* RELEVANCE OF DOCUMENTS IS DETERMINED BY A SIMILARITY (MATCHING) FUNCTION.

SEARCH STRATEGIES

- FULL-SEARCH (FS)
  - INEFFICIENT IF DB IS LARGE
  - EMULATED BY INVERTED INDEX SEARCH
- CLUSTER-BASED RETRIEVAL (CBR)

CLUSTER: A HOMOGENOUS GROUP OF DOCUMENTS THAT ARE MORE STRONGLY ASSOCIATED WITH EACH OTHER THAN WITH THOSE IN DIFFERENT GROUPS.

CLUSTERING: THE PROCESS OF FORMING THESE GROUPS

CLUSTERING HYPOTESIS:

"CLOSELY ASSOCIATED DOCUMENTS TEND TO BE RELEVANT TO THE SAME REQUEST"

CLUSTER BASED RETRIEVAL (CBR)

* QUERIES FIRST COMPARED WITH THE CLUSTERS. OR CLUSTER REPRESANTATIVES, CENTROIDS
* DETAILED QUERY-BY-DOCUMENT COMPARISION IS PERFORMED ONLY WITHIN SELECTED CLUSTERS.
* IT IS POSSIBBLE TO ARRANGE THE CLUSTERS IN A HIERARCHICAL STRUCTURE
* CBR FACILITATES BROWSING & EASY ACCES TO COMPLETE DOCUMET INFORMATION.
* TO INCREASE EFFICIENCY of CBR, DOCUMENT WITHIN A CLUSTER CAN BE STORED IN CLOSE PROXIMITY TO EACH OTHER WITHIN A DISK MEDIUM TO MINIMIZE I/O DELAY
* AN INDEX CAN BE BUILD TO SEARCH THE CENTROIDS FASTER.

CLUSTERING ALGORITHMS

CLASSIPICATION - I

- PARTITIONING TYPE (*)
- OVERLAPPING TYPE
- HIERARCHICAL TYPE

TO THE MANNER IN WHICH DOCUMENTS ARE DISTRIBUTED CLUSTERS

CLASSIFICATION - II

- SINGLE-PASS ALGORITHMS (*)
- ITERATIVE ALGORITHMS
- GRAPH THEORETICAL ALG.

ACCORDING TO THE CLUSTERING METHODOLOGY

(*) INDICATES METHODS USED IN C^3M

CONCEPT OF C^3M

* C^3M ALGORITHM IS PARTITIONING TYPE.
* CHOOSE A SET OF DOCUMENTS AS THE SEEDS AND TO ASSIGN ORDINARY (NONSEED) DOCUMENTS TO THE CLUSTERS INITIATED BY SEED DOCUMENT TO FORM CLUSTERS.
* COVER COEFFICIENT (CC) IS THE BASE CONCEPT OF C^3M CLUSTERING.

CC CONCEPT SERVES TO

1) IDENTIFY RELATIONSHIP AMONG DOCUMENTS OF A DATABASE BY USE OF A MATRIX
2) DETERMINE THE # OF CLUSTERS THAT WILL RESULT IN A DOC. DB.
3) SELECT CLUSTERS SEEDS USING CLUSTER SEED POWER
4) FORM CLUSTERS WITH RESPECTS TO C^2M, USING CONCEPT (1-3)
5) CORRELATE THE RELATIONSHIP BETWEEN CLUSTERING & INDEXING

CC CONCEPT

- NONHIERARCHICAL CLUSTERING
- SEED BASED
- C^3M, THE SEEDS MUST ATTRACT RELEVANT DOCUMENT ONTO ITSELF
- DOCUMENT RELATIONSHIP OF COVERAGE AND SIMILARITIES MUST BE DETERMINED IN MULTIDIMENSIONAL SPACE.
- THESE RELATIONSHIPS ARE REFLECTED IN THE C MATRIX WHOSE ELEMENTS CONVEY DOCUMENT/TERM COUPLING.

Definition: D MATRIX REPRESENTS THE DOCUMENT DATABASE. CC MATRIX, C, IS A DOCUMENT-BY-DOCUMENT MATRIX WHOSE ENTRIES, C_{ij} $(1 \leq i, j \leq m)$ INDICATE THE PROBABILITY OF SELECTING ANY TERM OF d_i FROM d_j

C MATRIX INDICATES THE RELATIONSHIP BETWEEN DOCUMENTS BASED ON A TWO-STAGE PROBABILITY EXPERIMENT. THE EXPERIMENT RANDOMLY SELECTS TERMS FROM DOCUMENTS IN TWO STAGES:

1) RANDOMLY CHOOSES A TERM t_k OF DOCUMENT d_i
2) CHOOSES THE SELECTED TERM t_k FROM DOCUMENT d_j

TO APPLY THE CC CONCEPT, THE ENTRIES OF THE D MATRIX, d_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$, MUST SATISFY THE FOLLOWING CONDITION

(1) EACH DOCUMENT MUST HAVE AT LEAST ONE TERM
(2) EACH TERM MUST APPEAR AT LEAST IN ONE DOCUMENT.

C_{ij}, ONE MUST FIRST SELECT AN ARBITRARY TERM OF d_i, SAY, t_k, AND USE THIS TERM TO TRY TO SELECT DOCUMENT d_j FROM THIS TERM, THAT IS, TO CHECK IF d_j CONTAINS t_k. THERE IS A TWO-STAGE EXPERIMENTS. EACH ROW OF THE C MATRIX SUMMARIZES THE RESULTS OF THIS TWO-STAGE EXPERIMENT.

Ex $A = \{\text{item is defective}\}$ $B_1 = \{\text{the item came from 1}\}$

$B_2 = \{\text{the item came from 2}\}$ $B_3 = \{\text{the item came from 3}\}$

Q= Suppose that one item is choosen from the stockpile and found to be defective. What is the probability that it was produced in factory 1?

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)} \qquad i = 1, 2, .., k$$

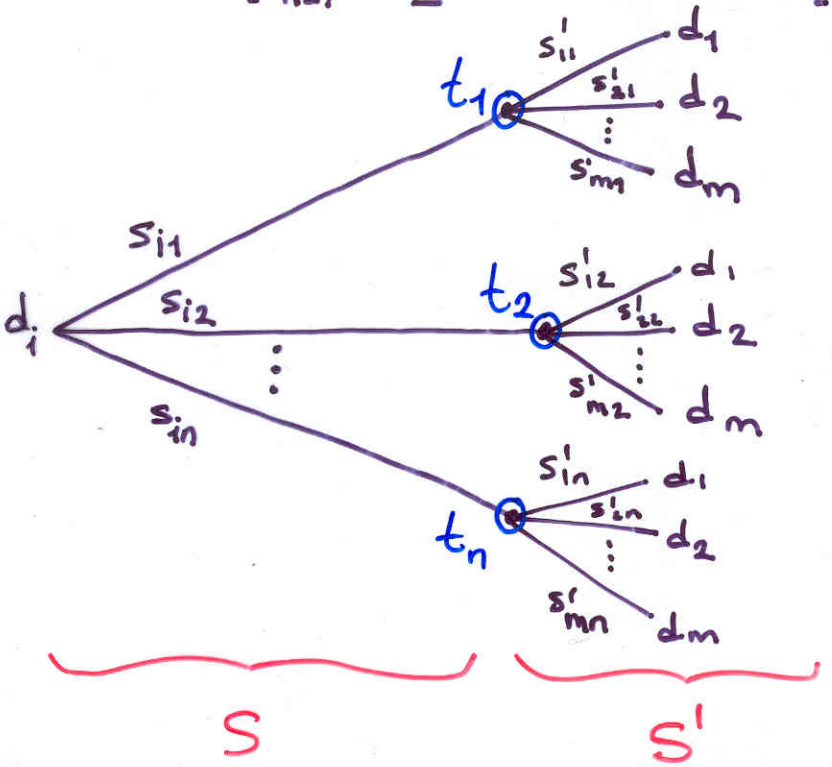
FIRST STAGE

S_{ik} INDICATE THE EVENT OF SELECTING t_k FROM d_i

SECOND STAGE

S'_{jk} INDICATE THE EVENT OF SELECTING d_j FROM t_k

$$S_{ik} = d_{ik} \times \left[\sum_{h=1}^{n} d_{ih} \right]^{-1}, \quad S'_{jk} = d_{jk} \times \left[\sum_{h=1}^{m} d_{hk} \right]^{-1}$$



- TO REACH FROM d_i TO d_j, THERE ARE n POSSIBLE WAYS.
- CHOOSE ONE OF THEM, t_k IS THE INTERMEDIATE STOP.
- IN ORDER TO REACH D_j WE MUST FOLLOW S'_{jk}.
- THE PROBABILITY OF REACHING d_j FROM d_i VIA t_k BECOMES

  $S_{ik} \times S'_{jk}$
- c_{ij} (THE PROBABILITY OF SELECTING A TERM OF d_i FROM d_j) IS EQUAL TO SUM OF THE PROBABILITIES OF INDIVIDUAL PATHS FROM d_i TO d_j.

$$D = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix}$$

$$C_{12} = \sum_{k=1}^{6} S_{1k} \times S'_{2k}$$

S_{1k} → probability of selecting t_k from d_i

S'_{2k} → probability of selecting d_j from t_k

$$= \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{4} + 0 \times 0 + 0 \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} + 0 \times 0$$

$$= 0.417$$

$$C_{ij} = \sum_{k=1}^{n} \alpha_i d_{ik} \cdot \beta_k = d_{jk}$$

$$\alpha_i = \frac{1}{\sum_{j=1}^{n} d_{ij}}$$ Row sum

$$\beta_k = \frac{1}{\sum_{i=1}^{m} d_{ij}}$$ Column sum

$$C_{ij} = \sum_{k=1}^{n} S_{ik} \times S'^{T}_{kj} = \sum_{k=1}^{n} (\text{probability of selecting } t_k \text{ from } d_i) \times (\text{probability of selecting } d_j \text{ from } t_k)$$

where $S'^{T}_{kj} = S'_{jk}$

PROPERTIES of C MATRIX

1) FOR $i \neq j$, $0 \leq c_{ij} \leq c_{ii}$ and $c_{ii} > 0$
2) $c_{i1} + c_{i2} + \cdots + c_{im} = 1$ (i.e., sum of row i is equal to 1)
3) If $c_{ij} = 0$, then $c_{ji} = 0$, and similarly, if $c_{ij} > 0$, then $c_{ji} > 0$; but in general, $c_{ij} \neq c_{ji}$ (nature of symmetry)
4) If none of the terms of d_i is used by the other documents, then $c_{ii} = 1$; otherwise, $c_{ii} < 1$.
5) $c_{ii} = c_{jj} = c_{ij} = c_{ji}$ iff d_i and d_j are identical

$$C = \begin{bmatrix} 0.417 & 0.417 & 0.000 & 0.083 & 0.083 \\ 0.313 & 0.438 & 0.000 & 0.063 & 0.188 \\ 0.000 & 0.000 & 0.333 & 0.333 & 0.333 \\ 0.083 & 0.083 & 0.111 & 0.361 & 0.361 \\ 0.063 & 0.188 & 0.083 & 0.271 & 0.369 \end{bmatrix}$$

$$c_{ij} = \begin{cases} \text{(Coupling of } d_i \text{ with } d_j\text{)} \\ \text{extent to which } d_i \text{ is covered by } d_j \text{ for } i \neq j \\ \text{extent to which } d_i \text{ is covered by itself for } i = j \\ \text{(decoupling of } d_i \text{ from the rest of the document)} \end{cases}$$

1) **Identical Documents:** coupling and decoupling is equal.

2) **Overlapping Documents:** Each document will cover itself more than any other. ($c_{ii} > c_{ij}$, $c_{jj} > c_{ji}$)

3) **A document is a subset of another document:** d_i subset of d_j. the extend to which d_i is covered by itself (c_{ii}) will be identical to the extent to which d_i is covered by d_j (c_{ij}) ($c_{jj} > c_{ii}$)

4) **Disjoint documents:** Since d_i and d_j do not have common terms, then they will not cover each other ($c_{ij} = c_{ji} = 0$)

$\delta_i = c_{ii}$ (decoupling coefficient of d_i) — How much the document is not related to other document.

$\psi_i = 1 - \delta_i$ (coupling coefficient of d_i)

$$\delta = \sum_{i=1}^{m} \delta_i / m = 1.945/5 = \underline{0.389}$$ Overall decoupling / Uniqueness

$$\psi = \sum_{i=1}^{m} \psi_i / m = 1 - 0.389 = \underline{0.611}$$ Overall Coupling

C' MATRIX

IT IS SIMILAR TO THE CONSTRUCTION OF THE C MATRIX. WE CAN CONSTRUCT A TERM-BY-TERM C' MATRIX OF SIZE n by n FOR INDEX TERMS.

$$c'_{ij} = \sum_{k=1}^{m} s'^{T}_{ik} \times s_{kj} = \sum_{k=1}^{m} (\text{probability of selecting } d_k \text{ from } t_i) \times (\text{probability of selecting } t_j \text{ from } d_k)$$

$$c'_{ij} = \sum_{k=1}^{m} \beta_i d_{ki} \times \alpha_k d_{kj}$$

Number-of-Cluster Hypothesis

THE NUMBER OF CLUSTERS WITHIN A DATABASE SHOULD BE HIGH IF INDIVIDUAL DOCUMENTS ARE DISSIMILIAR, AND LOW OTHERWISE

$$n_c = \sum_{i=1}^{m} \delta_i = \delta \times m \rightarrow 0.389 \times 5 = 1.945 \cong 2$$

- $n_c = 1$ iff all documents of D are identical
- For a binary D matrix, the minimum value of c_{ii} for $1 \leq i \leq m$ is $1/m$
- In a binary D matrix, $n_c > 1$ if we have at least two distinct document
- The value range of n_c is $1 \leq n_c \leq \min(m, n)$

CLUSTER SEED POWER

C^3M IS SEED-ORIENTED DOCUMENT-CLUSTERING METHODOLOGY.

- n_c DOCUMENT IS SELECTED AS CLUSTER SEED, AND NON-SEED DOCUMENTS ARE GROUPED AROUND THE SEEDS TO FORM CLUSTERS.
- SEED DOCUMENTS MUST BE WELL SEPERATED.
- SEED DOCUMENTS CANNOT BE TOO GENERAL OR TO SPECIFIC

$$P_i = \beta_i \times \Psi_i \times \sum_{j=1}^{n} d_{ij}$$

- $\sum_{j=1}^{n} d_{ij}$ → NORMALIZATION
- Ψ_i → INTRACLUSTER COHESION
- β_i → INTERCLUSTER DISPERTION

- THE FIRST n_c DOCUMENTS WITH THE HIGHEST SEED POWER ARE SELECTED AS THE SEED DOCUMENTS.

$P_2 = 0.985$

$P_5 = 0.957$

$P_1 = 0.729$

$P_4 = 0.692$

$P_3 = 0.222$

CONSTRUCTING CLUSTERS

FOR NON-SEED DOCUMENT d_1, $c_{12} = 0.417$ and $c_{15} = 0.083$.

SINCE $c_{12} > c_{15}$, d_1 WILL JOINT THE CLUSTER INITIATED BY d_2.

IF WE PROCEED IN THIS MANNER, GENERATED CLUSTERS WILL BE

$C_1 = \{d_1, d_2\}$ $C_2 = \{d_3, d_4, d_5\}$

CHARACTERISTICS of the C^3M ALGORITHMS

1) It enables us to estimate the number of clusters to be generated for a collection
2) We need to keep only the diagonal entries of the C matrix in determining the cluster seeds so that the space requirements of the algorithm is small
3) Document distribution within the clusters is rather uniform so that the cases with a few fat clusters or a lot of singleton are not countered.
4) Because, CC concept is used, the algorithm will be independent of the order in which documents clustered in the clustering process.
5) For average and the worst case complexity of the algorithms (single and multipass) are determined to be $O(m^2/\log m)$ for m documents.

INDEXING – CLUSTERING RELATIONSHIPS

n : number of (index) terms used in the description of doc.

m = total number of documents in DB

x_d = average number of terms used to describe a document (depth of indexing)

t_g = average number of documents described by a term (term generality)

$t = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij}$, total number of term assignment in D matrix (Nonzero entries in D matrix)

$$t_g = \frac{t}{n} \qquad x_d = \frac{t}{m}$$

$$n_c = \frac{m}{t_g} = \frac{n}{x_d} \qquad n_c = \frac{m \times n}{t} = \frac{t}{x_d \times t_g}$$

$$d_c = \frac{m}{n_c} = \frac{1}{\delta} = \frac{m}{(m/t_g)} = t_g$$

term generality

$$d_c' = \frac{n}{n_c'} = \frac{1}{\delta'} = \frac{n}{(n/x_d)} = x_d$$

Depth of Indexing

EXPERIMENTAL DESIGN AND EVALUATION

1) VALIDITY EXPERIMENTS TO TEST THE VALIDITY OF INDEXING-CLUSTERING RELATIONSHIP
2) USABILITY EXPERIMENTS TO VERIFY TIME EFFICIENCY OF C^3M
3) IR EXPERIMENTS TO MEASURE TO MEASURE THE PERFORMANCE OF C^3M

TODS214 —
- ACM Papers, March 1976 - 1984
- 214 document totaly
- Each document has, title, keywords, abstracts.
- Indexing generated automatically

INSPEC —
- Contains 12.684 documents
- Covering Comp. Science, Electrical Eng.

Characteristics of the D matrices for INSPEC database

D Matrix	min	max	n	t
D_{11}	1	5.851	14.573	412.255
D_{12}	2	5.851	7.435	405.117
D_{13}	2	3.000	7.431	387.811
D_{14} (*)	2	1.000	7.382	308.886
D_{15}	3	5.851	5.677	401.601
D_{16}	3	3.000	5.673	384.295
D_{17} (*)	3	1.000	5.624	305.370

(*) INDICATES THAT FOR THAT MATRIX m IS EQUAL TO 12.684

n : CARDINALITY OF THE INDEXING VOCABULARY

t : NUMBER OF NONZERO ENTRIES IN THE D MATRIX

RESULTS OF THE INDEXING-CLUSTERING RELATIONSHIP EXPERIMENTS FOR THE INSPEC DATABASE

Matrix	$\frac{m \times n}{t}$	n_{cb}	n_{cw}	t_g	d_{cb}	d_{cw}	x_d	d'_{cb}	d'_{cw}
D_{11}	448	439	475	28.29	28.89	26.70	32.50	33.20	30.68
D_{12}	233	227	275	54.49	55.88	46.12	31.94	32.75	27.04
D_{13}	243	238	294	52.19	53.29	43.14	30.58	31.22	25.28
D_{14}	303	294	364	41.84	43.14	34.84	24.36	25.11	20.28
D_{15}	179	175	218	70.74	72.48	58.18	31.66	32.44	26.04
D_{16}	187	183	233	67.74	69.31	54.44	30.30	31.02	24.35
D_{17}	234	227	289	54.30	55.88	43.88	24.08	24.78	19.46

# of clusters using CC — Avg. size of doc. clusters — Avg. size of term cluster.

USABILITY EXPERIMENTS

Time Efficiency:

- was coded in FORTRAN77
- Run on IBM 4381 Model 23
- UM/SP Operating System
- Execution times varies from 74 to 189 s

- —
- IBM 3083 BX environ.
- —
- 840 – 3416 s. (4.4 – 18.1) larger

Dmatrix	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}
ExecutionTime	189.00	126.00	114.00	85.00	105.00	95.00	74.000
t_{gs}	3.10	1.73	1.80	1.86	1.81	1.86	1.94
x_d	32.50	31.94	30.58	24.36	31.66	30.30	24.08
r	0.53	0.44	0.48	0.53	0.55	0.59	0.63

$r = (t_{gs} * x_d) / (e.t.)$ runtime is proportional to $(m \times x_d \times t_{gs})$

Validity of clustering structure

- Cluster validity is refered to objective ways of deciding whether a clustering structure represents the instrinsic character of a clustured data set.
- Given a query let a target cluster be defined as a cluster that contains at least one relevant document for the query
- For a valid clustering structure, n_t should be significantly less than the avarage number of target clusters under random clustering, n_{tr}
- If we preserve the same clustering structure and assign documents randomly to the clusters, the we obtain random clustering

Comparison of C^3M and Random Clustering in Terms of Avarage Number of Target Clusters for All Queries for the INSPEC Database

Dmatrix	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}
n_t	24.455	23.299	23.636	24.364	22.961	23.039	23.662
n_{tr}	30.807	29.580	29.715	30.379	28.769	29.005	29.765

$n_t < n_{tr}$

$n_t \geq n_{tr}$ — invalid

n_t : Avarage number of target clusters under given clustering structure

n_{tr} : Avarage # of target clusters, under random clustering

IR Experiments

Evaluation Measures for Retrieval Effectiveness:

- effectiveness of C³M by comparing its CBR performance with that of FS and with the CBR performance of other clustering algorithms used in the current IR literature.
- The effectiveness measures used in this study average precision for all queries.

T: Total # of relevant document retrieved for all queries.

Q: Total # of queries with no relevant documents.

- An effective system should have;
  - higher precision and T
  - lower Q

Term Weighting or Query Matching Function Selection

Term weighting basically has three components

1) Term Frequency Component (TFC)
   - b: binary weight
   - t: term frequency
   - n: Augmented normalized term frequency
2) Collection frequency Component (CFC)
   - x: no change
   - p: probabilistic inverse collection frequency factor
   - f: inverse document frequency
3) Normalization Component (NC)
   - x: no change
   - c: cosine normalization

$$\text{Similarity}(Q,D) = \sum_{k=1}^{n} w_{dk} \times w_{qk}$$

Term weighting Approaches Used in the Experiments

Abbrev.	TW1	TW2	TW3	TW4	TW5	TW6	TW7
Meaning	txc.txx	tfc.nfx	tfc.tfx	tfc.bfx	nfc.nfx	nfc.tfx	nfc.bfx

Generation of Cluster Centroids and Query Characteristics.

- The terms with the highest total number of occurences within the documents of a cluster are choosen as centroid terms.
- The maximum length (i.e., # of distinct terms) of centroids is provided as parameter.
- centroid length affects the efectiveness of CBR
- The weight of a centroid term is defined as its total number of occurences in the document of corresponding cluster.

Database	Centroid length (max x_c)	Avg. centroid length x_c	t_{gc}	$\%x_c$	$\%n$	$\%D$
TODS214	50	49.47	2.83	24	30	16
	100	94.81	3.06	46	54	30
	150	135.22	3.39	65	70	44
	200	163.69	3.51	79	81	53
INSPEC	250	237.09	12.96	51	60	27
	500	399.06	14.74	86	88	46

x_c : avarage # of distinct terms per centroids

t_{gc} : avarage # of centroids per term

$\%x_c$: percentage of the distinct cluster term used in the centroid

$\%n$: percentage of D matrix term that appear in at least one centroid

$\%D$: total size of centroid vectors as a percentage of the corresponding D matrix

Characteristics of the Query

Database	Number of Query	Avarage terms per query	Avarage relevant docs. per query	Total relevant document	Number of distinct docs. retrieved	Number of distinct terms for query definition
TODS 214	58	13.36	5.26	305	126	291
INSPEC	77	15.82	33.03	2543	1940	577

Effectiveness of CBR

INSPEC Database

- The first step in CBR is to determine the number of clusters (n_s) to be selected.
- Retrieving more clusters will be more effective, but also more expensive
- The increase in effectiveness would be expected to increase up to a certain n_s, and after this saturation point, the retrieval effectiveness remains same or improves very slowly
- for INSPEC database, this point is observed at $n_s = 50$ ($C \% 10$, $n_{cw} = 475$)