The Single Pass Clustering Method

S. Rieber and V. P. Marathe

Abstract

In information retrieval, several complex clustering methods exist which require extensive processing time and computer memory. A cheap clustering method is developed which requires only one pass over the document collection to generate clusters. The one-pass clustering method is investigated using the ADI collection of 82 documents and 35 queries which is available on-line in the SMART system. Clusters formed are not of uniform size; one or two early clusters are exceptionally large. Variation of the minimum correlation cutoff value is an adequate control for the number of clusters generated. The effect of document order on the clustering method is investigated, and the results are inconclusive. Overall, the single-pass clustering method is surprisingly effective and compares favorably with more complicated clustering methods.

1. Introduction

In information retrieval, a given item of information is often represented by an N-dimensional vector, each dimension referring to a different property of the item. Using matrix algebra, comparisons, classifications, and other processes can be performed on the information vector or on a long series of such vectors. Classifying the item, that is, placing it in a group of similar items, can save time later when it becomes necessary to refer to that item again. For example, at the supermarket it is much easier to find the aisle with breakfast cereals, and then to look for the corn

flakes, than it is to search for them on every shelf of every aisle.

The classification problem has two basic aspects:

1) classification already exists, and each new item is placed in that group to which it is the most similar;
2) no classification is assumed to exist and groups are formed on the basis of similarities between items. [1]

Both types of processes have advantages and disadvantages. If a classification already exists, the processing of new items is simple. However, the size of the groupings often becomes unbalanced in time with a resultant loss in efficiency. Merely finding the breakfast food section won't save much time if half the store is stocked with cereals. The second classification procedure of regrouping every time a few new items are added to the set is time-consuming. Many comparisons and decisions must be made to determine the optimum groups. It may be reasonable to include the same item in more than one group. But once the time and trouble have been taken to classify the items, any specific item is relatively easy to find.

In the context of automatic information retrieval, the items of information are documents and search requests, and the classification groups are called clusters. All the vectors in a cluster are averaged to obtain the cluster centroid, a vector representation of the entire group. Query vectors are correlated with the centroids. Only those centroids with sufficiently high correlations are expanded; that is each document in the cluster is correlated with the query and retrieved in order of higher correlation. This type of search is known as a cluster search or two-level serach. To contrast, a full search would correlate every document with the query and retrieve them in order of highest correlation. Cluster searching is a pro-

ven method for reducing search times hopefully without significant loss of relevant documents.

Many document grouping schemes representing both solutions are in existence. For example, library cataloging sorts publications into already existing classifications. On the other hand, much research is being done to implement the second classification process, that of forming clusters based only on the similarities between the items. In general these methods require extensive computer time and storage space to calculate correlations between every pair of documents. Rocchio's and Bonner's clustering methods [2], [3] are of this kind, since they require a processing time proportional to N^2, where N is the number of documents in the collection. A more efficient clustering method developed by Dattola [1] does not compute the document-document correlation matrix and reduces the processing time to one of order N Log N. A different, highly inexpensive method of clustering has been proposed; that of examining each document only once, and forming clusters in the process. Such a method is called the single-pass clustering method and is the subject of this paper.

The single-pass clustering method is appealing because it saves a significant amount of time over other clustering methods. To illustrate, assume that a clustering method requires 1 minute of computer time for 100 documents. Then to cluster 1,000,000 documents, an N^2 clustering method will take 170 years, an N Log N method 21 days, while the single-pass method would take only 7 days.

The single-pass clustering method also has a certain aesthetic appeal in that it represents a compromise between the two basic aspects of the classification problem. No initial classification exists, but as docu-

ments are processed, classifications are built up. A document is considered for inclusion into all existing clusters before it is allowed to start a cluster of its own. Classification groups are formed principally on the basis of the similarities of an item with an existing group rather than on the basis of similarities between items. However, once a document is accepted in an existing cluster, the cluster centroid is accordingly revised.

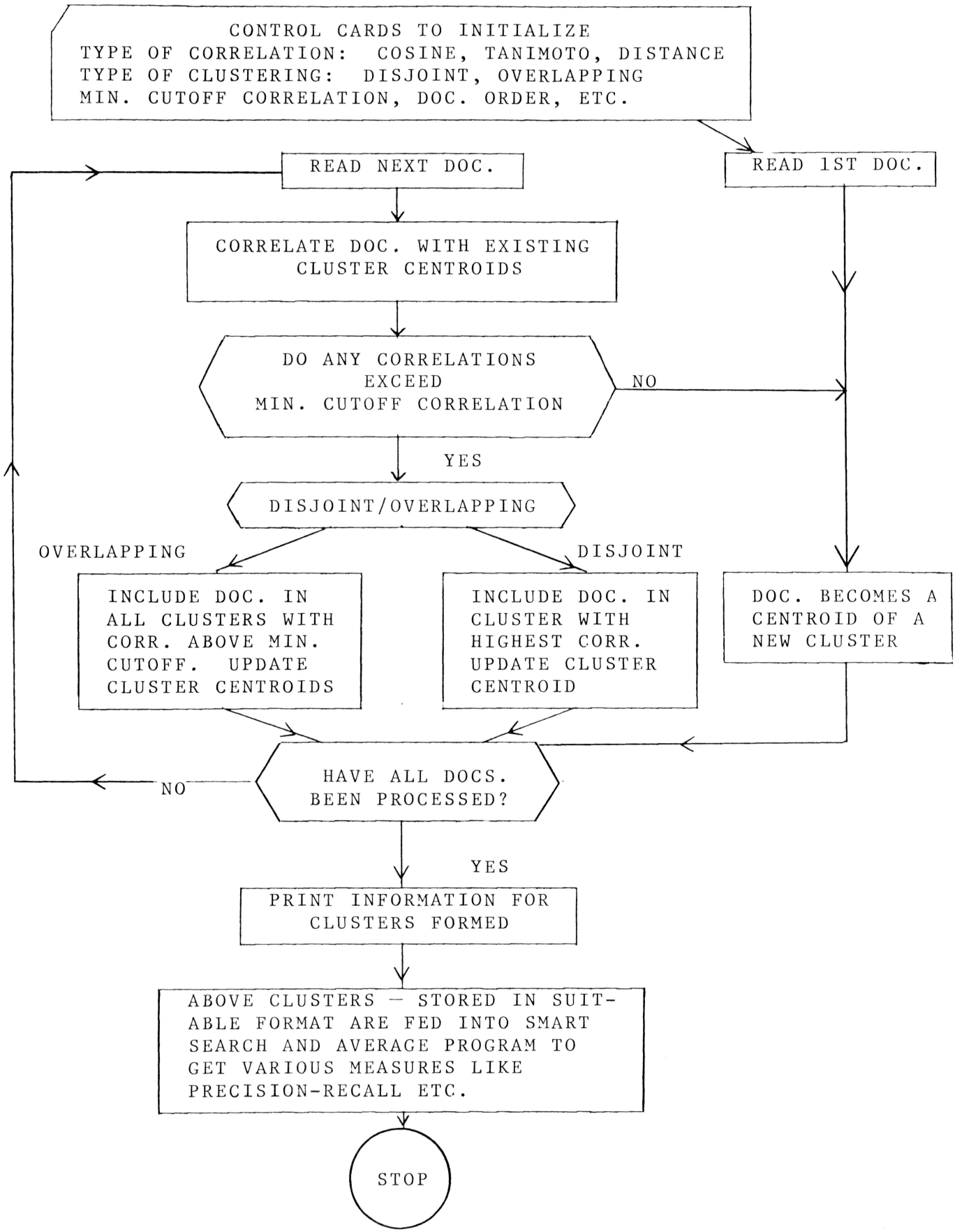
2. The Program

A FORTRAN program was written to implement the single pass clustering method. A flow diagram of the process is shown in Fig. 1.

The single-pass clustering method assigns the first document vector scanned as a cluster centroid. The next and succeeding documents are correlated with existing cluster centroids. If a minimum cutoff correlation is equaled or surpassed, the document is included in every such cluster and the cluster centroid is revised (overlapping), or the document is included only in that cluster with the highest correlation, assuming the minimum cutoff is again equaled or surpassed (disjoint). If the minimum correlation is not reached, a new cluster is formed with the document as centroid.

Two correlation methods — cosine and Tanimoto — can be used with the program. Cosine correlation normalizes all vectors to length 1.0 in N-dimensional space, N being the number of concept-dimensions in every document or query vector. The cosine of the angle between the two vectors is computed and used as a similarity measure.

$$\underline{\text{Cosine}} \qquad S_{d_1 d_2} = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| \cdot |\vec{d}_2|} = \frac{\sum_{i=1}^{N} d_1^i \cdot d_2^i}{\left(\sum_{i=1}^{N} (d_1^i)^2 \cdot \sum_{i=1}^{N} (d_2^i)^2 \right)^{\frac{1}{2}}}$$



Single Pass Clustering Methods

Fig. 1

Tanimoto's correlation also performs vector multiplication to obtain a similarity measure. The denominator is a normalizing term to prevent dense document vectors from being assigned disproportionately high correlations.

$$\underline{\text{Tanimoto}} \qquad S_{d_1 d_2} = \frac{\vec{d}_1 . \vec{d}_2}{\vec{d}_1 . \vec{d}_1 + \vec{d}_2 . \vec{d}_2 - \vec{d}_1 . \vec{d}_2}$$

$$S_{d_1 d_2} = \frac{\sum_{i=1}^{N} d_1^i \, d_2^i}{\sum_{i=1}^{N} (d_1^i)^2 + \sum_{i=1}^{N} (d_2^i)^2 - \sum_{i=1}^{N} (d_1^i d_2^i)}$$

The program control parameters are as follows:

1. Type of clustering — overlapping or disjoint.
2. Correlation Method — cosine or Tanimoto.
3. Minimum Concept Weight for Centroid — To prevent the dilution of the cluster centroid, concept weights of less than this value are set to zero. All runs were made at .005.
4. Concept bound — the highest concept number of the document collection.
5. Mult — Concept weights of the normalized centroid vector in the single-pass program are less than 1.0. To prevent most of these concepts from disappearing when converting to integer (fixed point) values for the SMART system, all centroid weights are multiplied by MULT. All runs were made with MULT equal to 100.
6. Cutoff — The value of the minimum correlation cutoff, below which a document will not be included in a cluster.

Six control cards cause the above quantities to be introduced into the program, and also determine whether the output will be printed or punched or

both. Four of the cards are name cards and place the proper SMART control cards at the beginning of the output to facilitate SMART evaluation.

3. Investigation and Results

Using the 82 document, 35 query American Documentation Institute collection filed on-line in the SMART CDS (Cornell Data Set), the following investigations of the single-pass clustering method were undertaken:

1) correlation comparison: cosine correlation, forward order — comparison of clusters formed by the two different correlation methods over a range of correlation cutoff values;

2) Disjoint-overlapping comparison: cosine correlation, forward order — comparison of disjoint and overlapping clusters over a range of correlation cutoff values;

3) Variation of document order: cosine correlation, fixed correlation cutoff — comparison of both overlapping and disjoint clusters formed with three different initial document orders;

4) SMART evaluation: SMART evaluation of the 6 sets of clusters formed in part 3 to compare overlapping and disjoint methods as well as the three document orders;

   a) full search of the document collection and evaluation for comparison to single-pass runs;

   b) SMART evaluation of clusters formed by Dattola's method, for comparison;

   c) SMART evaluation of Dattola's clusters formed from initial single pass clusters.

A) Correlation Comparison

For cosine correlations, correlation cutoff values ranging from .10 to .30 were investigated. The number of clusters formed varied from 5 at

.10 to 31 at .30 with 14 clusters formed at a correlation cutoff of .20. Tanimoto correlation cutoff was varied from .02 to .30. The number of clusters formed varied from 7 at .02 to 81 at .30 with 15 clusters formed at a cutoff of .04. All clustering was disjoint using forward document order. Appendix 1 contains tablesand graphs of the results of all runs performed for this phase of the investigation.

For every correlation method, the relationship between the number of clusters generated and the correlation cutoff values is smooth and not unmanageably steep. Suitable variation of the correlation cutoff value can therefore be used to roughly control the number of clusters that will be formed for a given document order.

A striking fact about cosine and Tanimoto clusters is the large variation between sizes of clusters formed, as shown in Fig. 2. The largest clusters are usually among the first 2 or 3 clusters, no doubt because they are formed early and have a chance to inspect nearly the entire document collection. Tanimoto clusters seem to be slightly more evenly distributed than cosine clusters. There is no apparent relationship between documents included in a given cosine cluster and documents included in a given Tanimoto cluster.

B) Disjoint-Overlapping Comparison

For cosine correlation, forward document order, the correlation cutoff was varied from .10 to .30, and both overlapping and disjoint clusters were formed. At a cutoff of .20, 14 disjoint and 17 overlapping clusters were formed. The average document is only included in one cluster for the disjoint method but is included in 4.8 clusters for the overlapping method. Apparently the dilution of the overlapping clusters resulted in the forma-

Cosine .20 Correlation Cutoff		Tanimoto .04 Correlation Cutoff	
Number of Clusters	Number of Documents in each Cluster	Number of Clusters	Number of documents in each Cluster
2	20 or 21	1	23
4	5,6, or 7	3	8,9,10, or 11
2	3 or 4	1	5,6, or 7
6	1 or 2	5	3 or 4
		5	1 or 2
14 Total Clusters	5.9 Average Documents per Cluster	15 Total Clusters	5.5 Average Documents per Cluster

Cluster Size Variations

Fig. 2

tion of three more clusters than the disjoint method. Any relationship between the document composition of a given overlap cluster and that of a given disjoint cluster is submerged by the large amount of overlap. The results of the SMART evaluation runs made on disjoint and overlapping clusters will be discussed in that section. Tables and graphs for all runs appear in Appendix 1.

C) Variation of Document Order

Three different initial document orders were used for this investigation — forward, reverse, and middle. The forward document order lists the documents sequentially, while the reverse order has them backwards. Middle order introduces the documents in an inside-out order; for example, documents 1, 2, 3, 4, 5, 6 in middle order would be 4, 3, 5, 2, 6, 1. Using a fixed cosine correlation cutoff of .20, both disjoint and overlapping clusters were formed with the three different initial document orders. Fig. 3 shows the number of clusters formed for each situation.

From this data, it can be tentatively concluded that a 10% or 20% variation in the number of clusters formed from one initial document order to another will not be unusual.

D) SMART Evaluation

Precision and recall data were averaged over all disjoint and all overlapping runs to obtain a measure of effectiveness reasonably independent of document order. Precision-recall graphs of the following runs are plotted in Appendix 2.

1. Full search
2. Dattola's cluster search

FOL (Forward Overlap) - 17 clusters
MOL (Middle Overlap) - 14 clusters
ROL (Reverse Overlap) - 19 clusters
FDJ (Forward Disjoint)- 14 clusters
MDJ (Middle Disjoint) - 14 clusters
RDJ (Reverse Disjoint)- 19 clusters
Average Cluster Size - 16.2 clusters

$$\text{Square Rt. of Std. DeV.} = \left(\frac{\sum_{i=1}^{N} (x_i - x_{AV})^2}{N-1} \right)^{1/2} = 2.07 \text{ clusters.}$$

Variation in Cluster Size
for Different Document Orders

Fig. 3

3. Single-pass disjoint cluster search
4. Single-pass overlapping cluster search

The full search curve is naturally more effective than other curves. The single-pass disjoint cluster search is the next best search. Dattola's cluster search appears higher at both ends of the graph than the single-pass overlapping cluster search, but loses some ground in the middle to high-recall region. Fig. 4 shows for each run, the index of rank recall plus log precision, another method of system evaluation, and reveals nearly the same results.

Using one set of disjoint and one set of overlapping clusters as initial clusters, Dattola's clustering method was evaluated. Precision-recall graphs of these results are found in graph 2 of Appendix 2. The disjoint initial clusters seem to give a higher precision at low recall values but a lower precision at high recall values than the overlapping initial clusters. For comparison Dattola's previous run with unknown initial clusters is also plotted and falls in the middle.

In Appendix 2 graphs 3 and 4, the precision-recall curves of the disjoint and overlapping clusters used as initial clusters for Dattola's method are plotted. Also Dattola's results using these clusters as initial clusters are plotted to determine if any significant improvement in clustering has occurred. For the disjoint clusters no improvement occurrs. Dattola's clustering slightly reduces the effectiveness of the initial clusters. For the overlapping clusters, Dattola's method increases precision at the low recall end but reduces it at the high recall end of the graph, a significant improvement.

It is felt that the apparently high recall and precision results ob-

Search Type	Rank Recall plus Log Precision
Full	.6323
Average Disjoint Cluster	.5359
Average Overlapping Cluster	.5086
Dattola's Cluster	.4888

A Parameter for Comparing Various Methods

Fig. 4

tained by the single-pass method relative to Dattola's method is in part due to the expansion of one or two very large clusters, providing a search of over half the document collection. Four more runs were made to reduce the number of clusters expanded so that a more legitimate comparison between the two methods could be made. Although a reduction in the number of clusters expanded occurrs, no appreciable effect is observed on the performance measures, probably due to the continued presence of those extremely large clusters.

For each run, the average documents checked per query are tabulated in Fig. 5. Most single-pass runs checked far more documents than Dattola's runs, revealing that a single-pass cluster search is not as efficient as a Dattola cluster search in terms of search time. Recall ceilings, however, naturally favor the system which looks at the most documents.

Global comparisons of all evaluation runs are tabulated in Table 2 of Appendix 2. Normalized precision, normalized recall, rank recall, log precision, rank recall plus log precision, and recall ceiling are included.

4. Conclusions

Over the 82 document 35 query ADI collection, the single-pass clustering method as described here is effective. Because the document collection is so small, these promising results cannot be conclusive. More extensive evaluation runs with larger document collections should be conducted before allowing the single-pass clustering method a permanent place in the information retrieval world.

The wide gap in the number of clusters formed for different input orders contributes to a growing suspicion that the single-pass method may

be very order-dependent. The three orders used were not truly random. The middle order, for example, is a composite of both the forward and reverse orders as shown in Fig. 6. Further investigations of the single-pass clustering method should include evaluation with random initial document orders to determine more precisely the standard variation of the number of clusters formed at a given correlation cutoff.

Disjoint clusters are more effective than overlapping clusters. Reducing the figure of 4.3, the number of overlapping clusters to which the average document belongs, would certainly improve the overlapping feature. It is recommended that the overlapping algorithm be revised so that a document is included in the cluster with which it correlates the highest (above cutoff) and only in one or two other clusters if the correlation is within the value of a small parameter, epsilon, of its highest correlation.

The number of clusters which will be formed can be effectively controlled by the establishment of a range of values for the correlation cutoff.

Perhaps the biggest pitfall of the single-pass method is the formation of one or two excessively large clusters. It is these large clusters which cause too many documents to be checked and reduces the efficiency of the cluster search. To prevent large clusters from forming, it is recommended that the single-pass clustering program be extended so that, as cluster size increases, the cosine correlation cutoff value slides up through a sequence of values. As the number of items in a cluster increases, it becomes more difficult for a document to assimilate into that cluster. Such a program feature would reduce the large variation in the sizes of the clusters.

Finally the possibility of more than one pass through the document collection to determine the optimum cosine correlation cutoff should be con-

Run Description	Average Documents Checked per Query	Recall Ceiling
Dattola (MOL initial clusters)	27.3	.47
Dattola (MDJ initial clusters)	23.8	.53
RDJ (Reverse disjoint)	25.5	.58
FDJ (Forward disjoint)	33.0	.66
MDJ (Middle disjoint)	39.0	.66
MOL (Middle overlapping)	45.6	.66

Comparison of Average Documents
for Various Methods

Fig. 5

Forward Order	4		5		6	
Middle Order	4	3	5	2	6	1
Reverse Order		3		2		1

Relation in Three Document
Orders Taken — An Illustration

Fig. 6

sidered. For large document collections a small representative subset could be reprocessed until an optimum correlation cutoff is determined.

References

[1] Dattola, R. T., A Fast Algorithm for Automatic Classification, Report ISR-14 to the National Science Foundation, Section V, Department of Computer Science, Cornell University, 1968.

[2] Bonner, R. E., "On Some Clustering Techniques", IBM Journal of Research and Development, Vol. 8, No. 1, January 1964.

[3] Rocchio, J. J., Jr., Document Retrieval Systems — Optimization and Evaluation, Report ISR-10 to the National Science Foundation, Harvard Computation Laboratory, Cambridge, Massachusetts, March 1966.

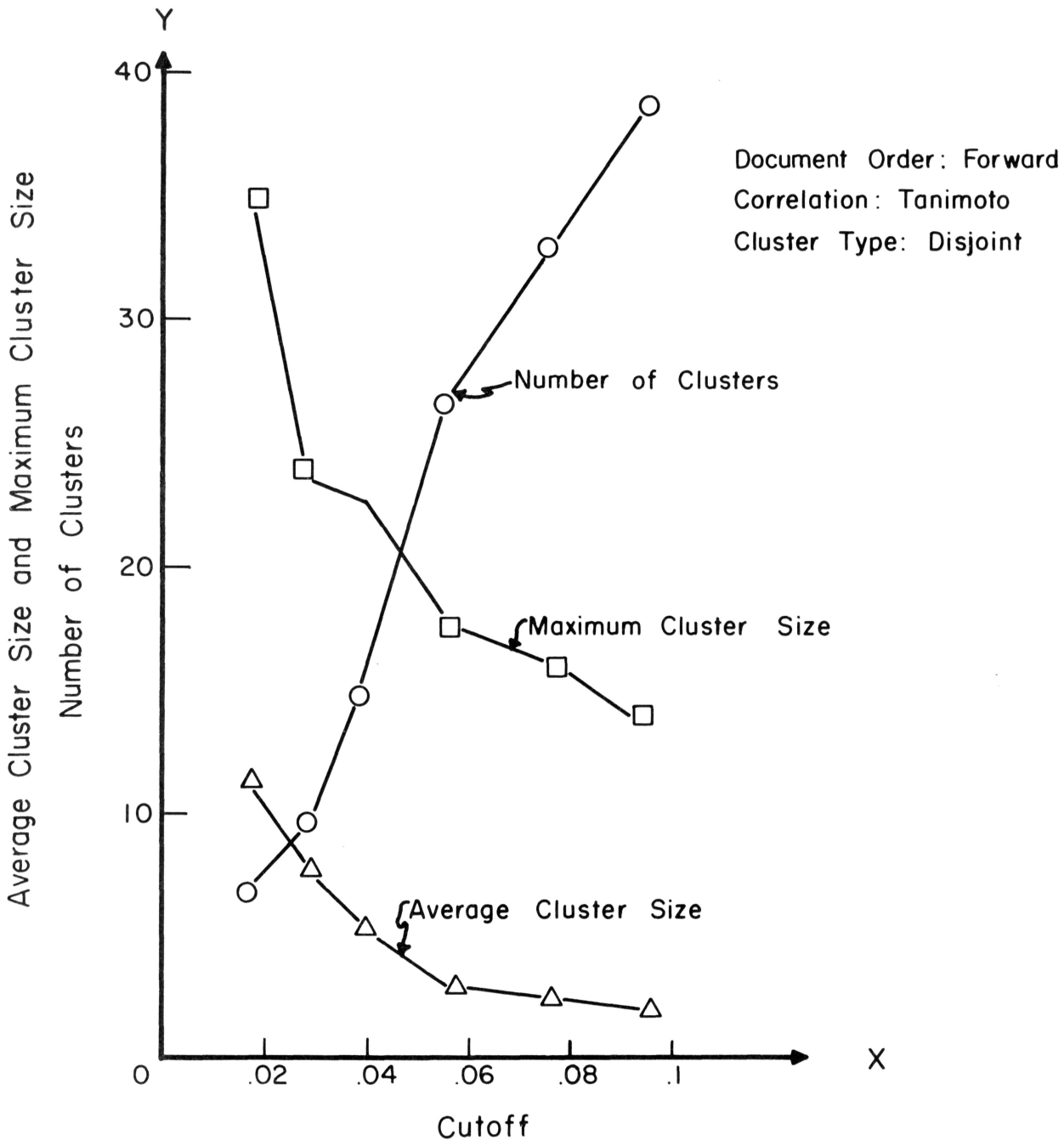
Appendix 1

Cluster Size Information

Cutoff	Number of Clusters	Average Size of Clusters	Maximum Size of Clusters	Minimum Size of Clusters
0.02	7	11.7	35	1
0.03	10	8.2	24	2
0.04	15	5.5	23	1
0.06	27	3.0	18	1
0.08	33	2.5	17	1
0.10	39	2.1	14	1
0.20	>50	1.2	4	1
0.30	>50	1.0	1	1

Cluster Size Observations for Forward Document Order,
Tanimoto Correlation and Disjoint Clusters

Table 1



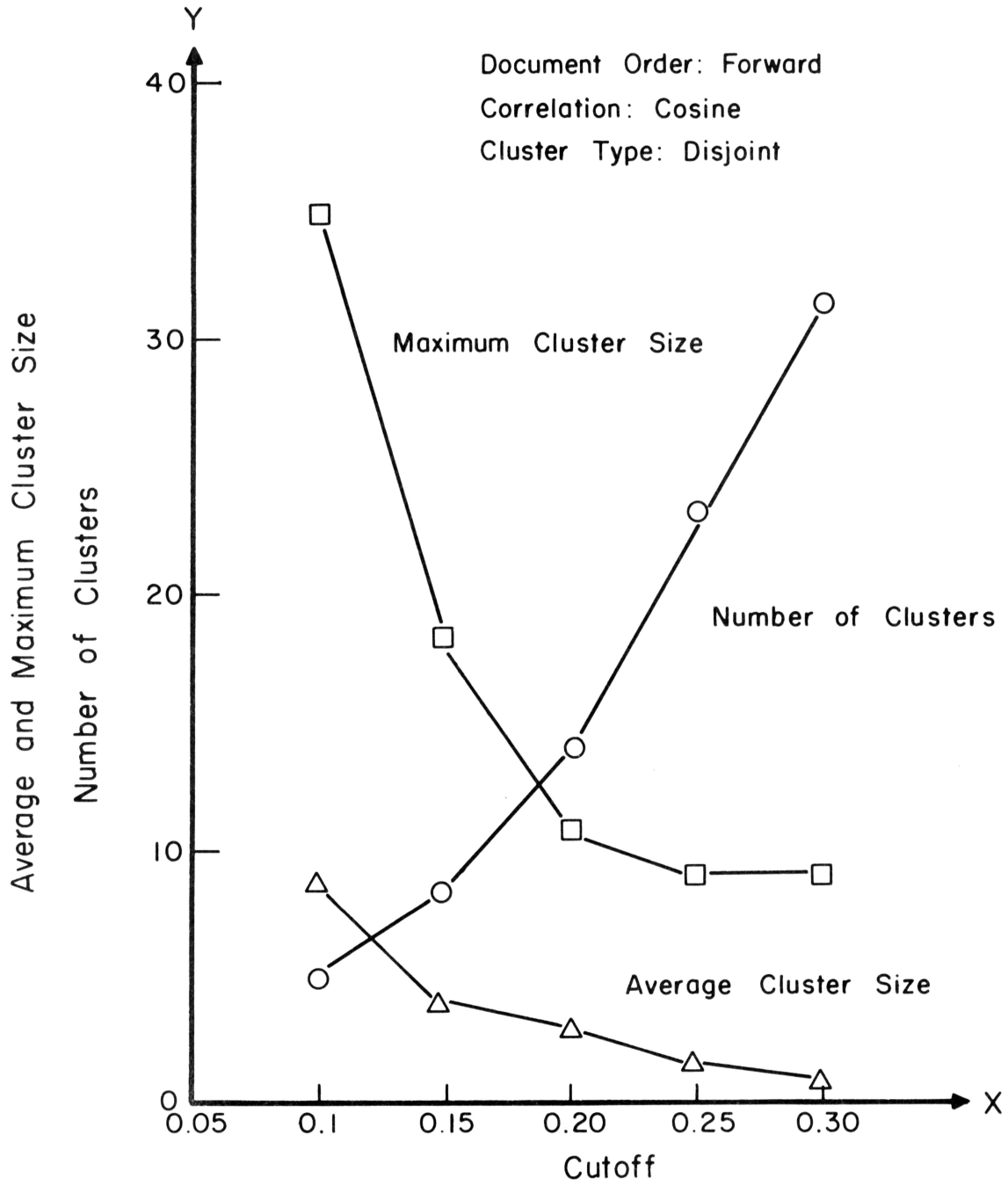
Recall-Precision Graph Corresponding to Table 1

Graph 1

Cutoff	Number of Clusters	Cluster Size		
		Average	Maximum	Minimum
0.10	5	16.4	70	1
0.15	9	9.1	37	1
0.20	14	5.9	21	1
0.25	23	3.6	19	1
0.30	31	2.6	19	1

Cluster Size Observations for Forward Document Order,
Cosine Correlation and Disjoint Clusters

Table 2



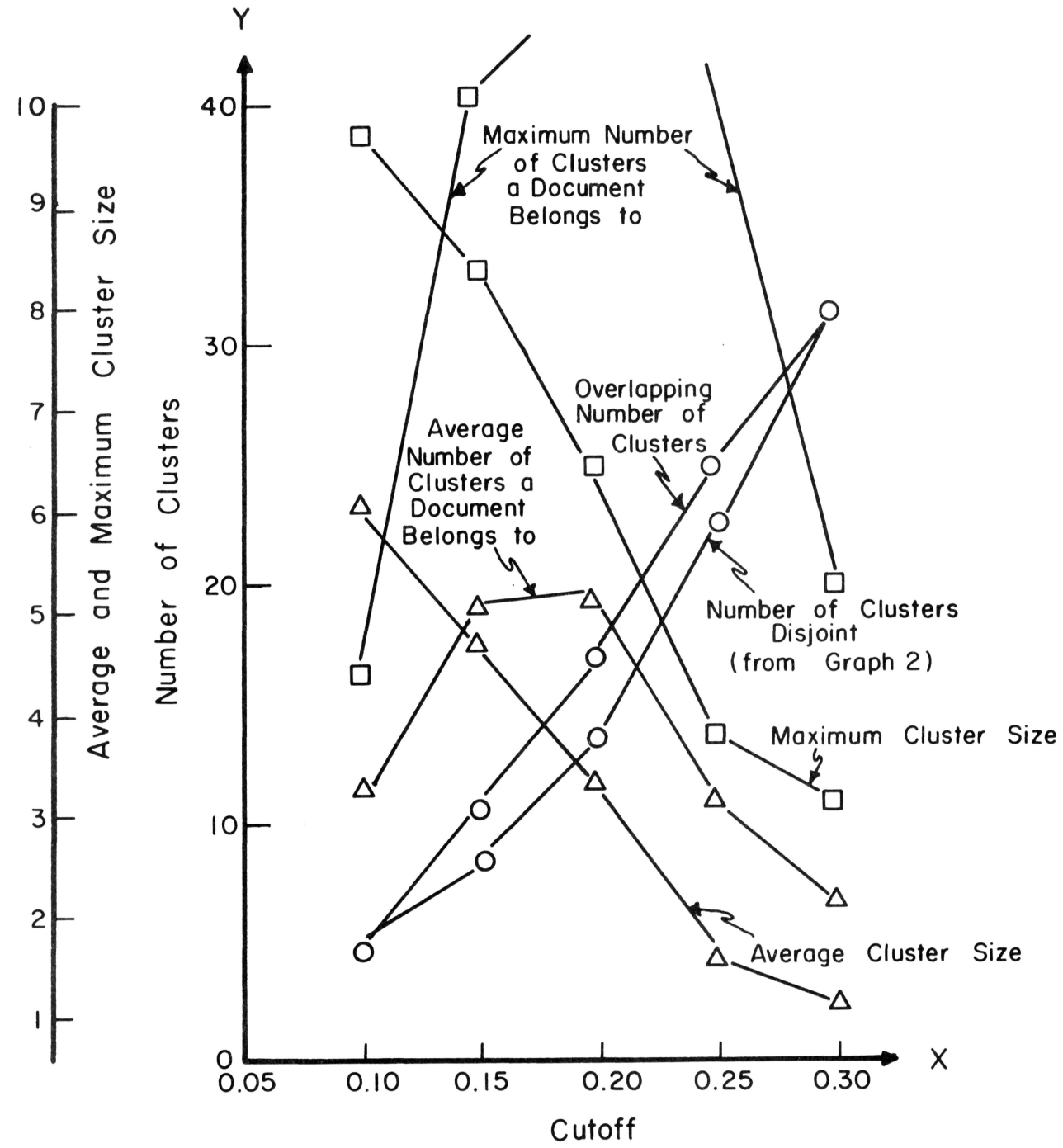
Recall-Precision Graph Corresponding to Table 2

Graph 2

Cutoff	Number of Clusters	Cluster Size			Number of Clusters A Document Belongs To		
		Average	Maximum	Minimum	Average	Maximum	Minimum
0.10	5	48.0	78	1	2.9	4	1
0.15	11	35.3	66	1	4.7	10	1
0.20	17	23.2	50	1	4.8	12	1
0.25	25	9.2	27	1	2.8	12	1
0.30	31	4.4	21	1	1.7	5	1

Cluster Size Observations for Forward Document Order,
Cosine Correlation and Overlapping Clusters

Table 3



Recall-Precision Graph Corresponding to Table 3

Graph 3

Document Order	Cluster Type	Number of Clusters	Cluster Size			Number of Clusters A Document Belongs To		
			Average	Maximum	Minimum	Average	Maximum	Minimum
Forward	Disjoint	14	5.9	21	1	1	1	1
	Overlapping	17	23.2	50	1	4.8	12	1
Reverse	Disjoint	19	4.3	24	1	1	1	1
	Overlapping	19	23.5	45	1	5.4	13	1
Middle	Disjoint	14	5.9	29	1	1	1	1
	Overlapping	14	16.4	47	1	2.8	7	1

Cluster Size Observations for Cosine Correlation
at Minimum Correlation Cutoff of 0.20

Table 4

Appendix 2

Clustering Evaluation

Recall	Full Search	Dattola's Clusters	Dattola's Clusters Using Initial Cluster As		Single Pass Clusters	
			MOL[1]	MDJ[2]	Average* Overlapping	Average* Disjoint
0.0	0.6356	0.6152	0.5468	0.6192	0.5999	0.6253
0.05	0.6356	0.6152	0.5468	0.6192	0.5999	0.6253
0.10	0.6356	0.6142	0.5468	0.6192	0.5981	0.6251
0.15	0.6142	0.5900	0.5332	0.6008	0.5784	0.6036
0.20	0.5690	0.5580	0.4955	0.5522	0.5269	0.5564
0.25	0.5350	0.4957	0.4587	0.5051	0.4951	0.5268
0.30	0.4939	0.4686	0.4213	0.4736	0.4555	0.4784
0.35	0.4604	0.4192	0.3930	0.4361	0.3931	0.4438
0.40	0.4582	0.4068	0.3870	0.4342	0.3864	0.4417
0.45	0.4482	0.3952	0.3527	0.4085	0.3734	0.4248
0.50	0.4458	0.3932	0.3508	0.4064	0.3725	0.4235
0.55	0.3424	0.2235	0.2656	0.2825	0.2620	0.3063
0.60	0.3399	0.2235	0.2644	0.2816	0.2611	0.3048
0.65	0.3172	0.2144	0.2573	0.2652	0.2452	0.2869
0.70	0.2453	0.1809	0.1819	0.1765	0.1854	0.2117
0.75	0.2438	0.1809	0.1819	0.1765	0.1839	0.2117
0.80	0.2342	0.1748	0.1726	0.1576	0.1734	0.2064
0.85	0.2173	0.1585	0.1600	0.1445	0.1567	0.1903
0.90	0.2053	0.1585	0.1572	0.1208	0.1471	0.1773
0.95	0.2053	0.1585	0.1572	0.1208	0.1471	0.1773
1.00	0.2053	0.1585	0.1572	0.1208	0.1471	0.1773

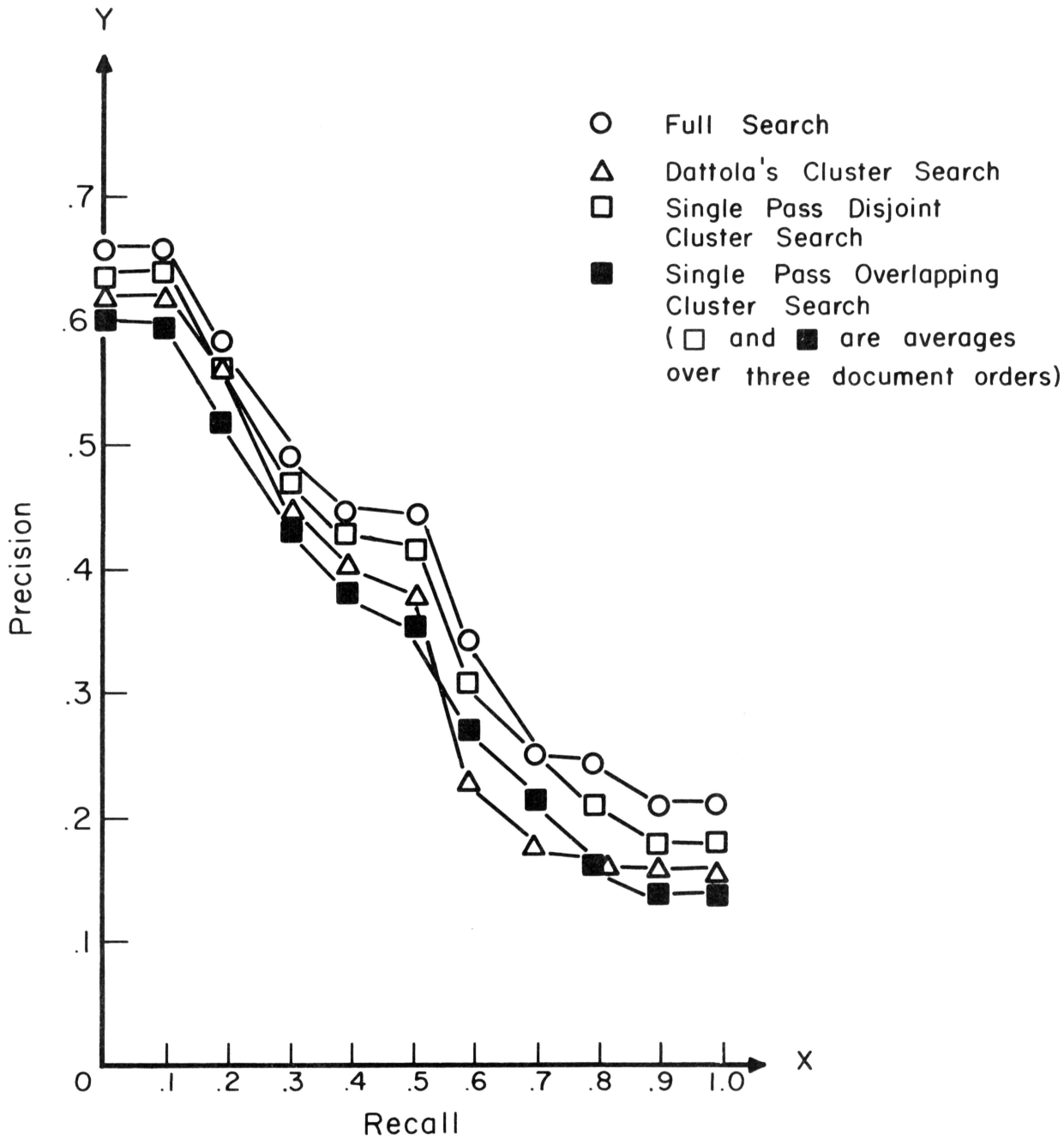
1. MOL ≡ Single pass clusters with Cosine Correlation, Cutoff 0.2, Document order — Middle, Cluster type — Overlapping.

2. MDJ ≡ Same as above, except Cluster type — Disjoint.

* These Average Precisions are taken over three Document orders

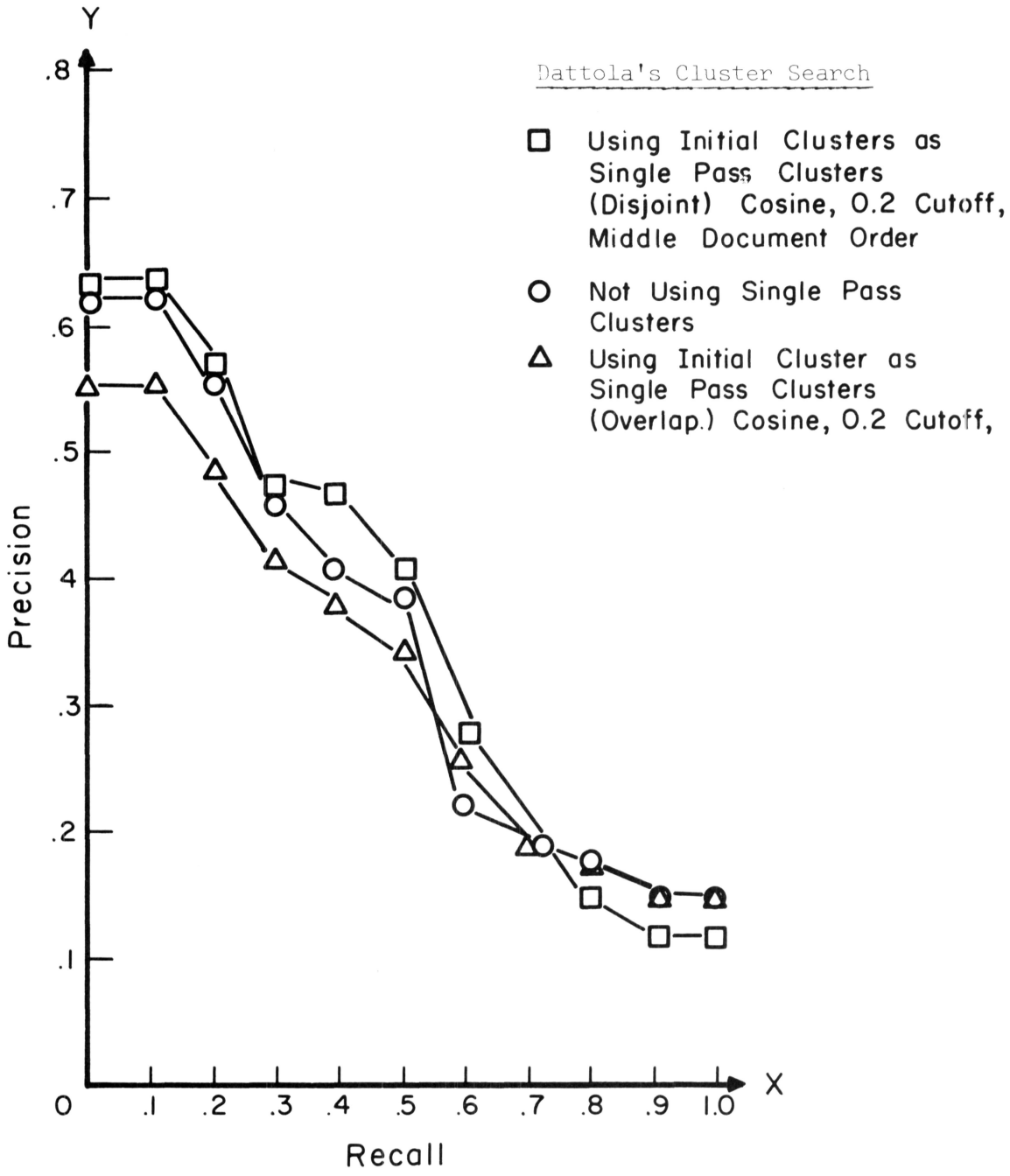
Recall-Precision Tables

Table 1



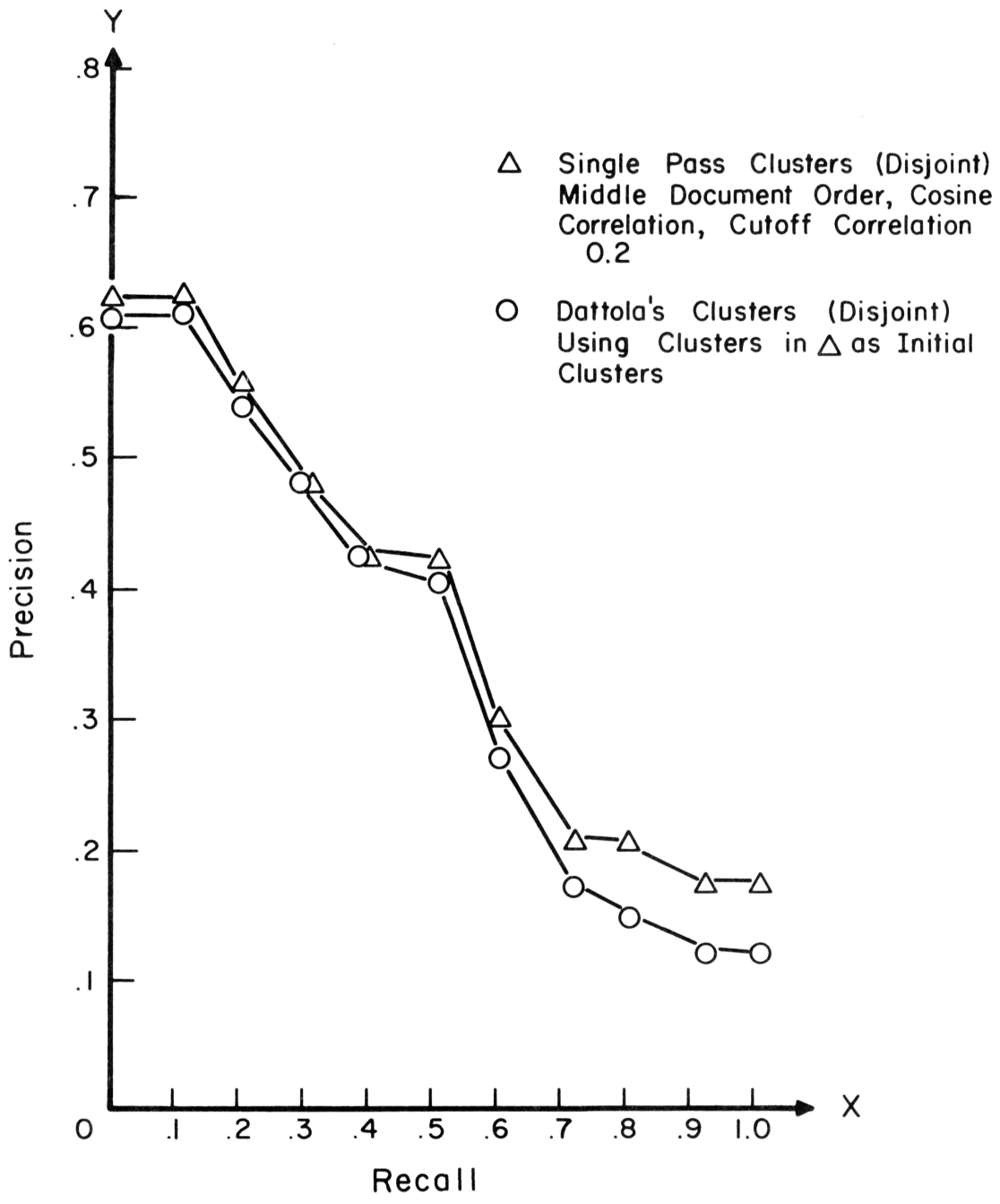
Recall-Precision Graph Corresponding to Table 1

Graph 1



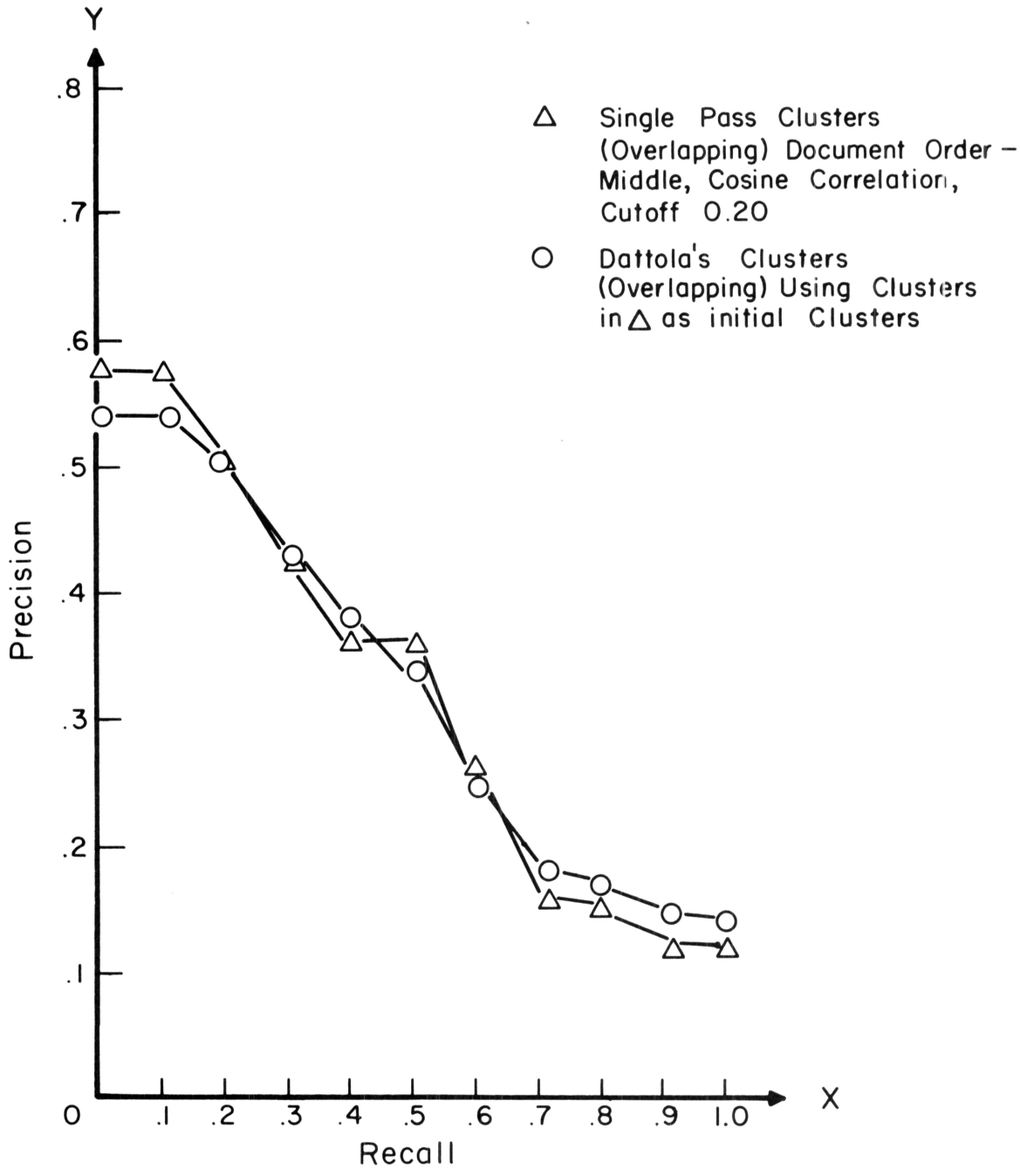
Recall-Precision Graph Corresponding to Table 1

Graph 2



Recall-Precision Graph Corresponding to Table 1

Graph 3



Recall-Precision Graph Corresponding to Table 1

Graph 4

Type of Search / Measures	Full	Dattola's Clusters (12)	Dattola's Clusters Using Initial Clusters As		Single Pass Clusters*					
			MOL* (11)	MDJ* (11)	FDJ (14)	MDJ (14)	RDJ (19)	FOL (17)	MOL (14)	ROL (19)
Normalized Recall	0.7946	0.5207	0.4630	0.5011	0.6075	0.5960	0.5497	0.6176	0.5842	0.5698
Normalized Precision	0.6055	0.4592	0.4198	0.4583	0.5218	0.5104	0.4900	0.5169	0.4784	0.4772
Rank Recall	0.2500	0.1737	0.1756	0.1507	0.2200	0.2109	0.2028	0.1962	0.1649	0.1770
Log Precision	0.3823	0.3151	0.3286	0.3351	0.3637	0.3512	0.3454	0.3446	0.3234	0.3204
Rank Recall + Log Precision	0.6323	0.4888	0.5042	0.4858	0.5837	0.5621	0.5482	0.5408	0.4883	0.4974
Recall Ceiling (Average)	—	—	0.49	0.53	0.66	0.66	0.58	0.69	0.66	.60

*Single pass clusters with Cosine Correlation, Cutoff 0.2 — The three letter name indicates —
First letter for Document order — F: Forward, M: Middle, R: Reverse
Next two letters for Cluster type — DJ: Disjoint, OL: Overlapping
For example: MDJ ≡ Middle Document order, Disjoint clusters
The number in () indicates number of clusters before search.

Global Evaluation Measures

Table 2