Paper Presentation

Hierarchical Document Clustering Using Frequent Itemsets

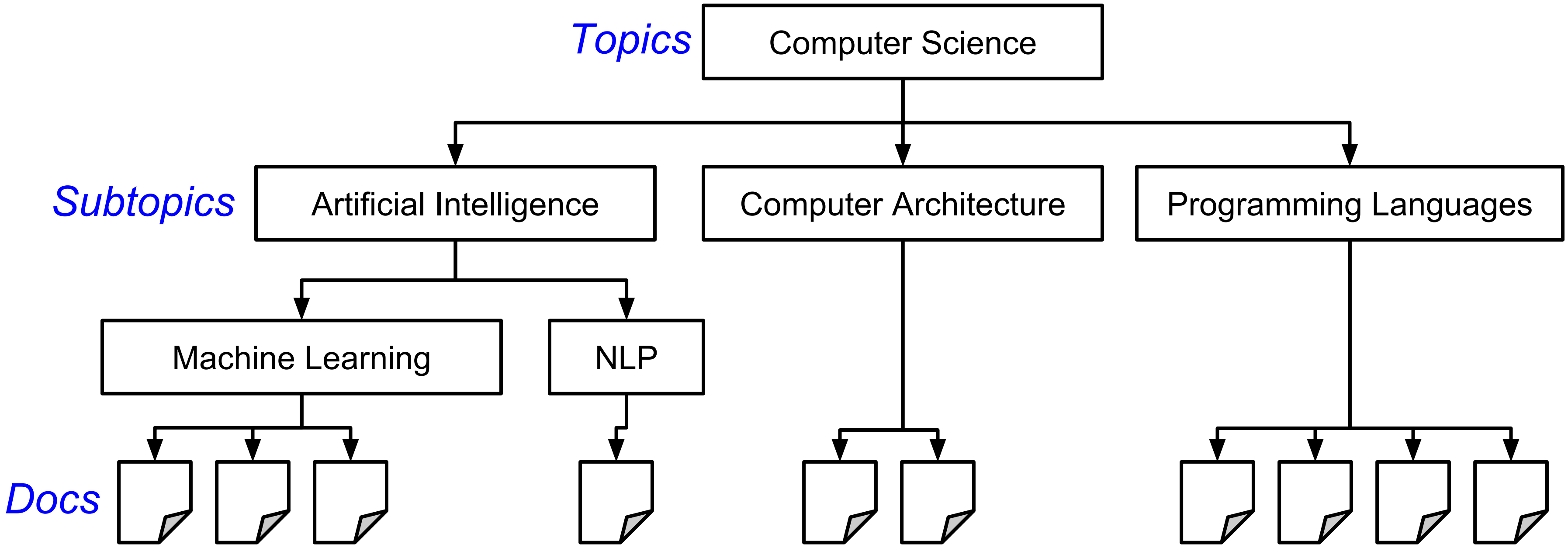
(B. Fung, K. Wang, and M. Ester. *SDM* 2003)

Presenters: Deyuan Guo, Elaheh Sadredini
CS@UVa. March 28, 2016

Outline

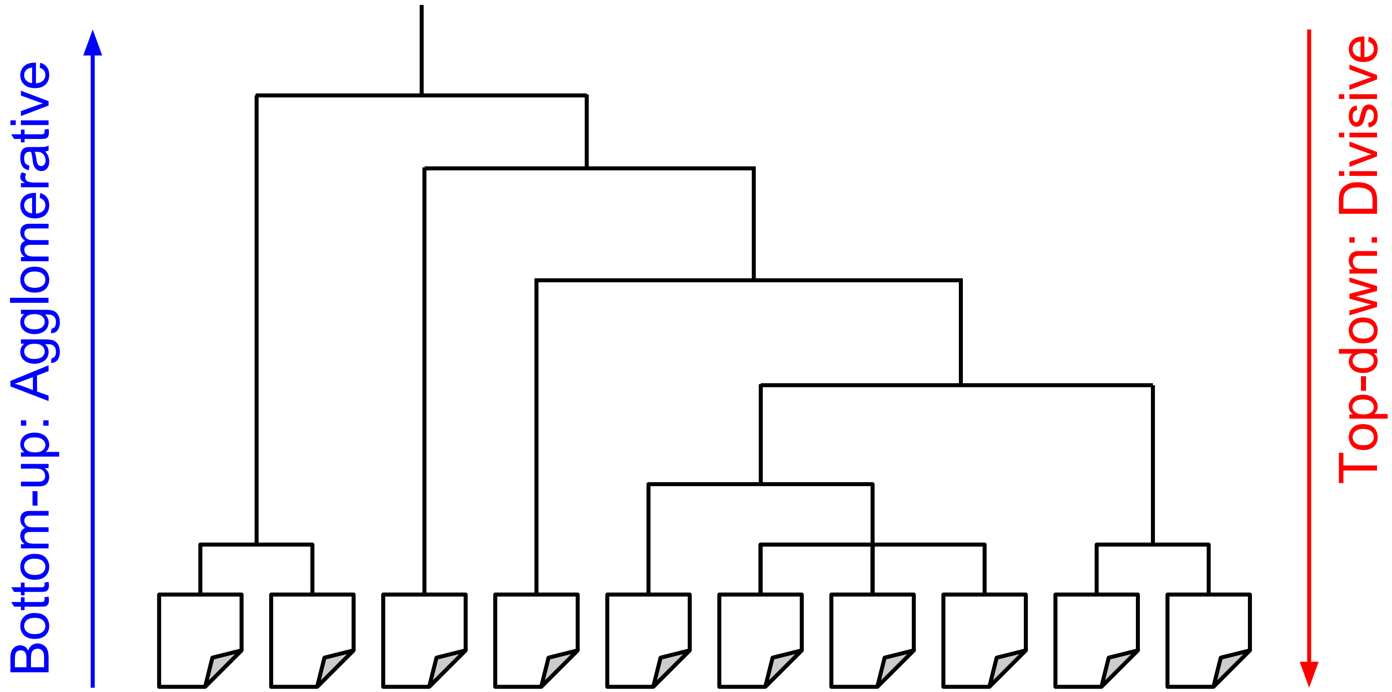
- Background
- The Frequent Itemset-based Hierarchical Clustering (FIHC) Approach
- Experimental Results

- Hierarchical Document Clustering



- Challenges in hierarchical document clustering
  - High dimensionality
  - High volume of data
  - Consistently high clustering quality
  - Meaningful cluster description

- Two types of hierarchical document clustering



UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (Kaufman and Rousseeuw, 1990)

Bisecting K-means (Steinbach, Karypis, and Kumar, 2000)

Background

- Frequent itemset-based approaches
  - Previous work: Document clustering using **frequent itemsets**, by Wang et al., 1999. (No hierarchy)
  - Previous work: Hierarchical Frequent Term-based Clustering (**HFTC**), by Beil, Ester, and Xu, 2002. (Greedy heuristic, not scalable)

- Today's topic: Frequent Itemset-based Hierarchical Clustering (**FIHC**)
  - Cluster-centered: Measure the similarity of clusters directly using frequent itemsets
  - Overcome many challenges: High dimensionality; Scalability; Meaningful cluster description; Accuracy; etc.

Preprocessing

- Stopword removal
- Stemming
- Vector model (TF × IDF)

Doc 1: apple = 5, boy = 2, cat = 7

Doc 2: apple = 4, window = 3

Doc 3: boy = 3, cat = 1, window = 5

	apple	**boy**	**cat**	**window**
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	0	3	1	5

FIHC - Basic Definitions

Global frequent itemset

Global support

Global frequent item

Cluster frequent item

Cluster support

	apple	boy	cat	window
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	0	3	1	5

FIHC - Basic Definitions

Global frequent itemset

Global support = 60%

Global frequent item

Cluster frequent item

Cluster support

	apple	boy	cat	window
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	0	3	1	5

FIHC - Basic Definitions

Global frequent itemset

Global support = 60%

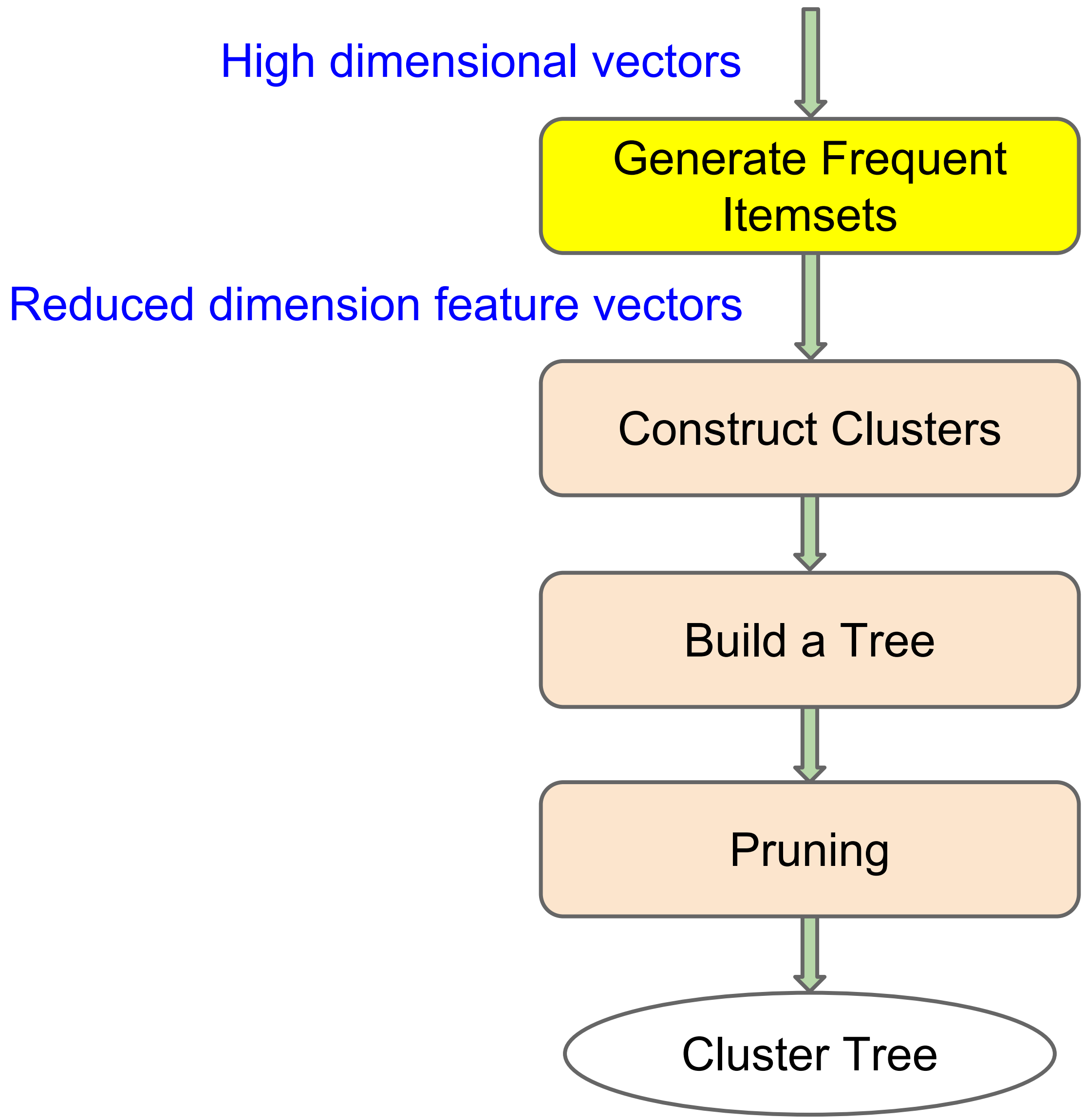
Global frequent item = boy or cat

Cluster frequent item

Cluster support

	apple	boy	cat	window
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	0	3	1	5

FIHC Algorithm Overview



FIHC - Generate Frequent Itemsets

	apple	boy	cat	window
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	1	3	1	5
Doc 4	8	0	2	0
Doc 5	5	0	0	3

Minimum support = 60%

Frequent itemsets: {apple}, {cat}, {window}, {apple, window}

FIHC - Generate Frequent Itemsets

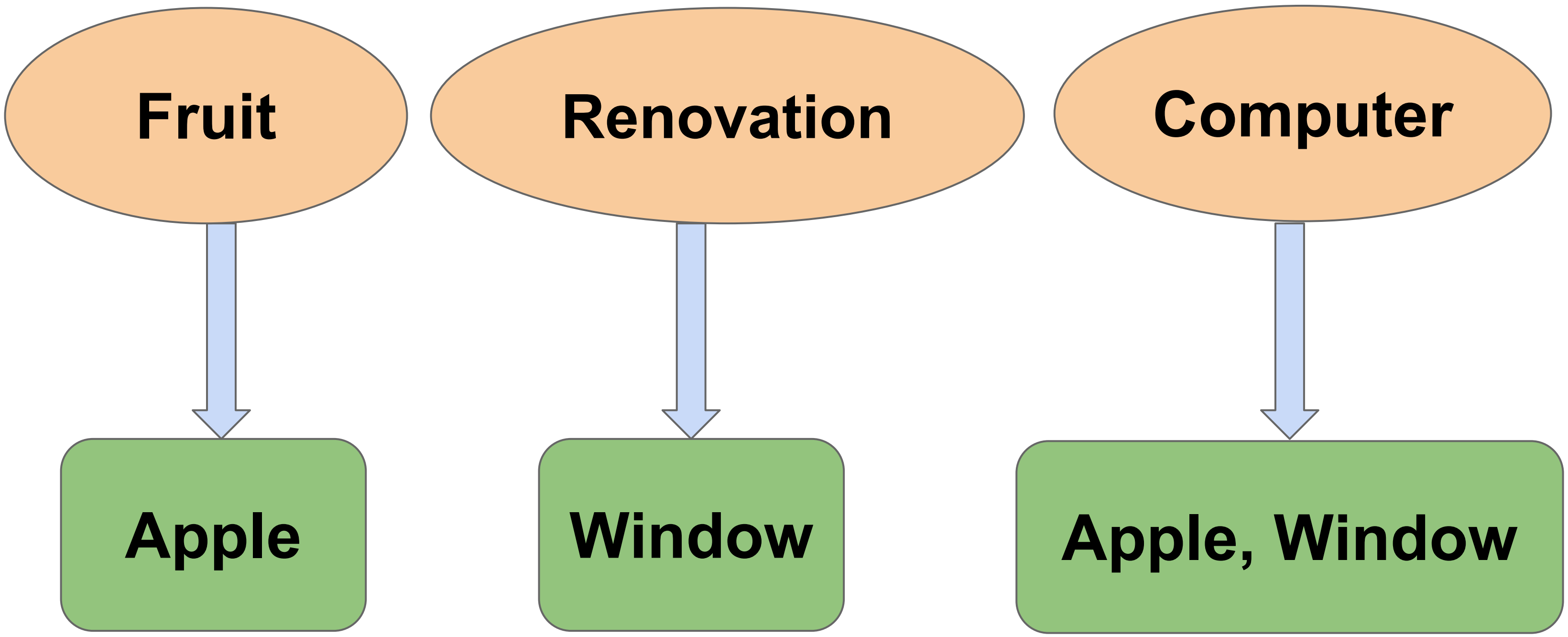
Reduce dimension → Improves efficiency and scalability

	apple	boy	cat	window
Doc 1	5	2	7	0
Doc 2	4	0	0	3
Doc 3	0	3	1	5
Doc 4	8	0	2	0
Doc 5	5	0	0	3

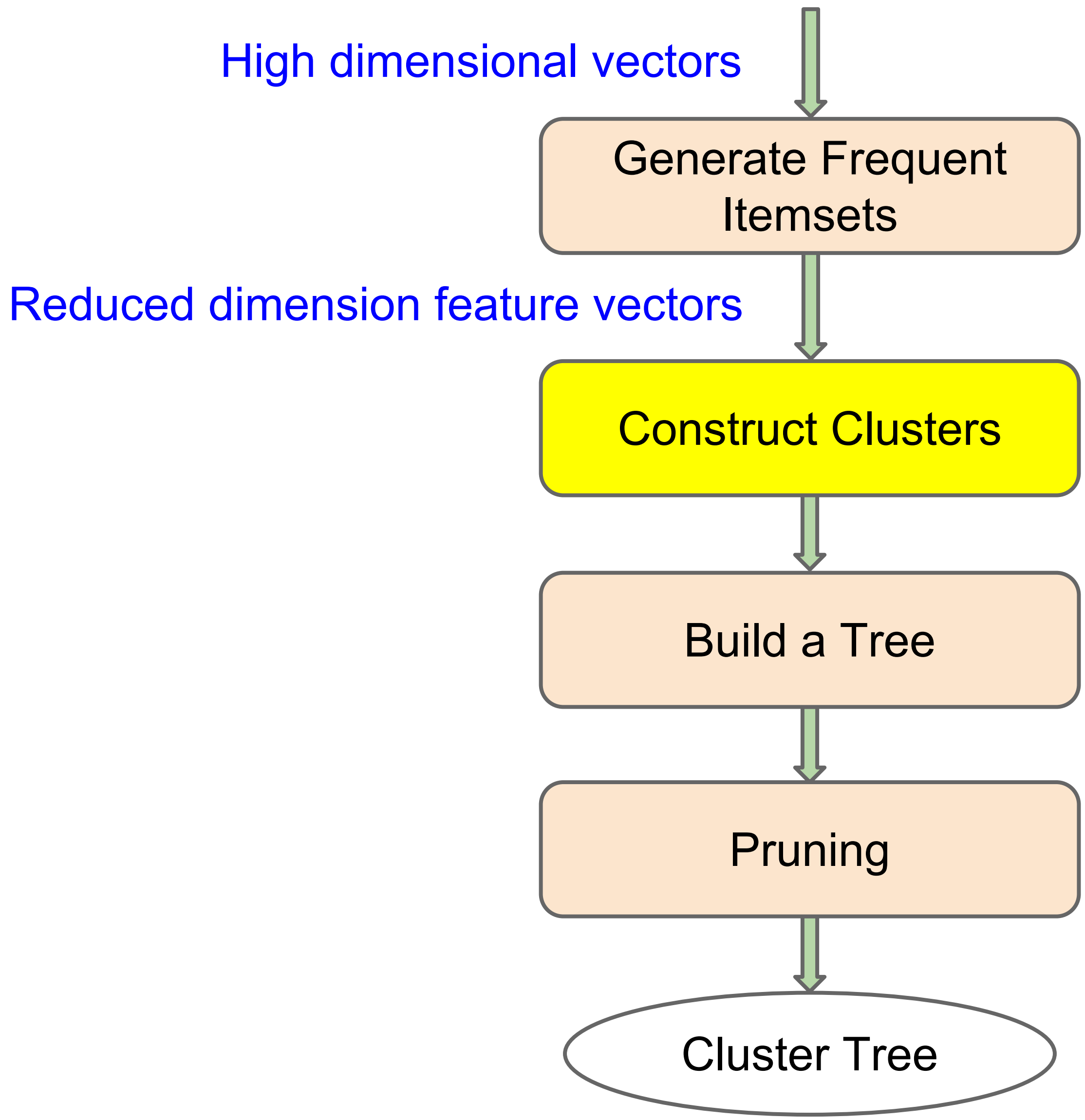
Minimum support = 60%

Frequent itemsets: {apple}, {cat}, {window}, {apple, window}

FIHC - Why Frequent Itemset Mining?

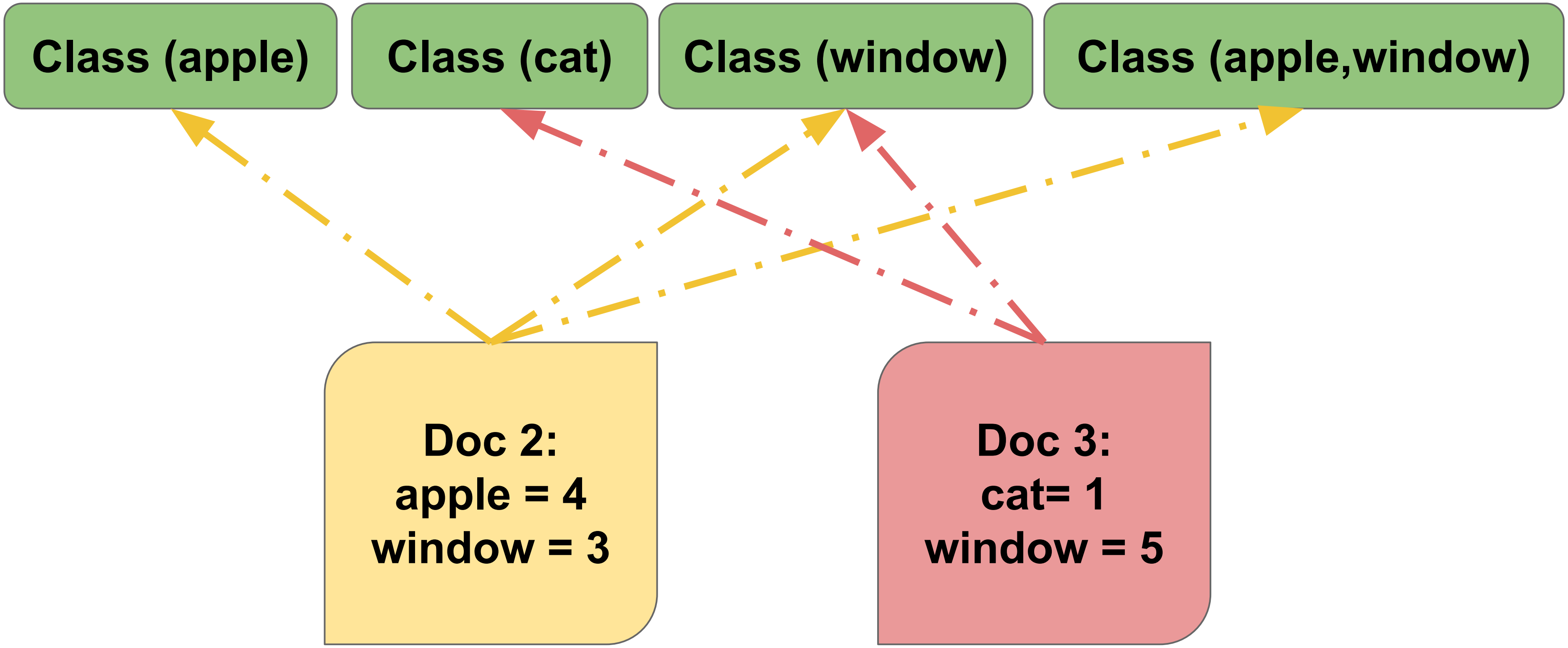


FIHC Algorithm Overview



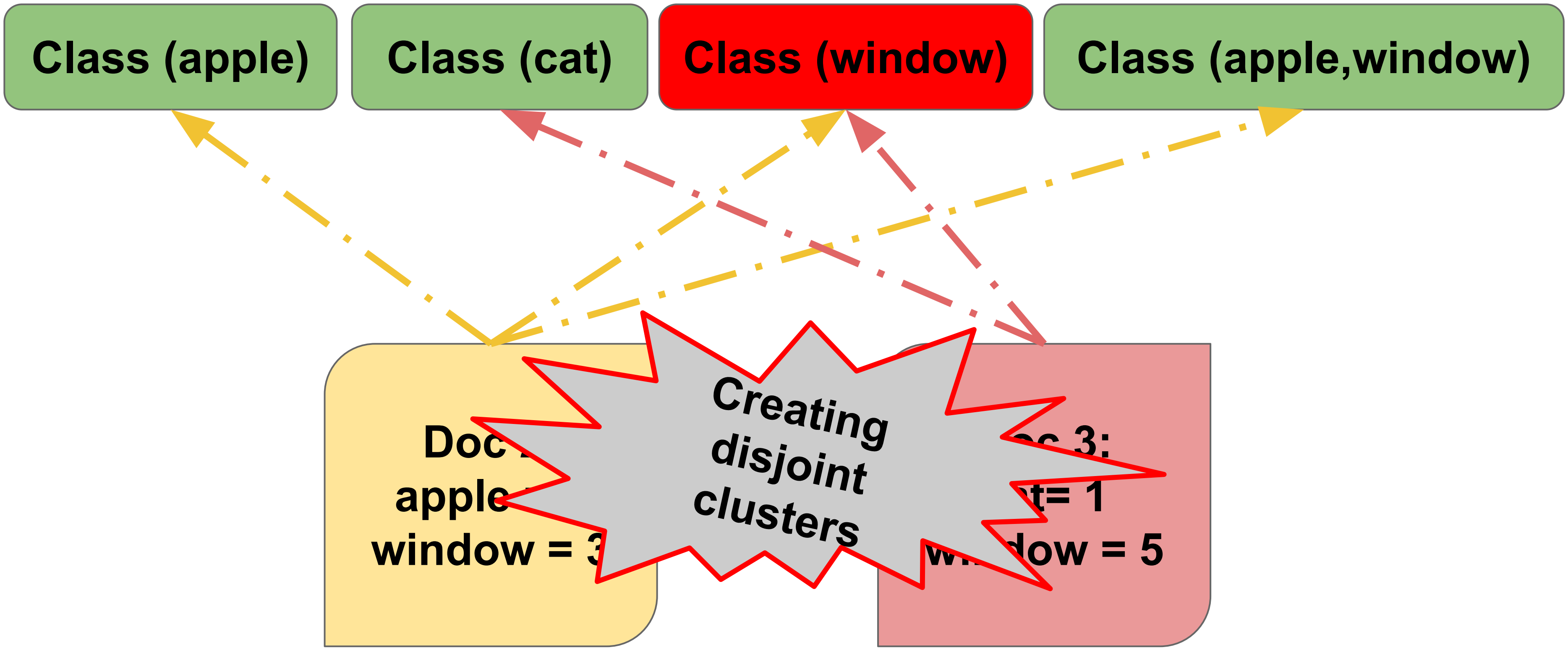
FIHC - Construct Initial Clusters

Frequent itemsets: {apple}, {cat}, {window}, {apple, window}



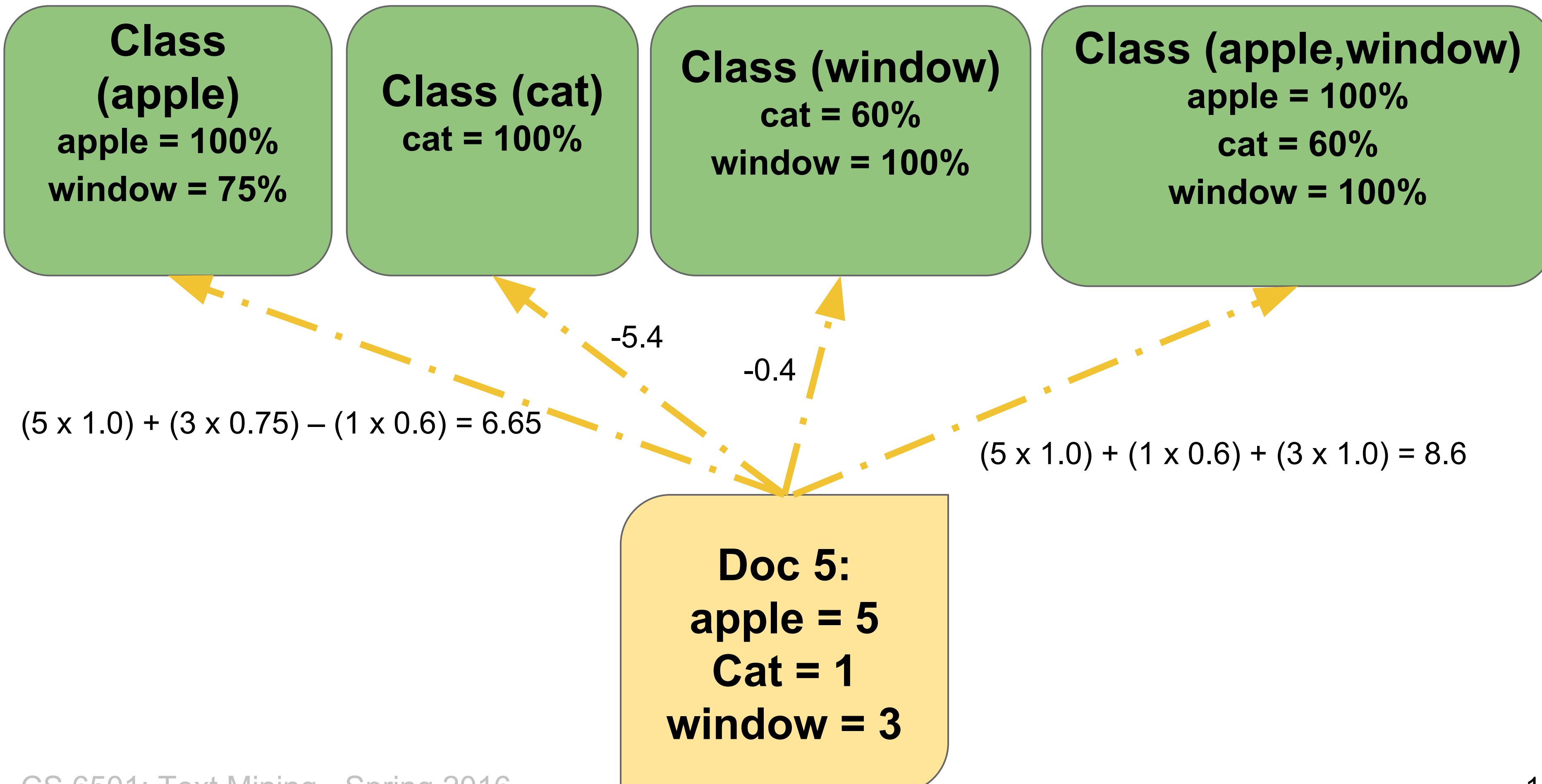
FIHC - Making Clusters Disjoint

Frequent itemsets: {apple}, {cat}, {window}, {apple, window}

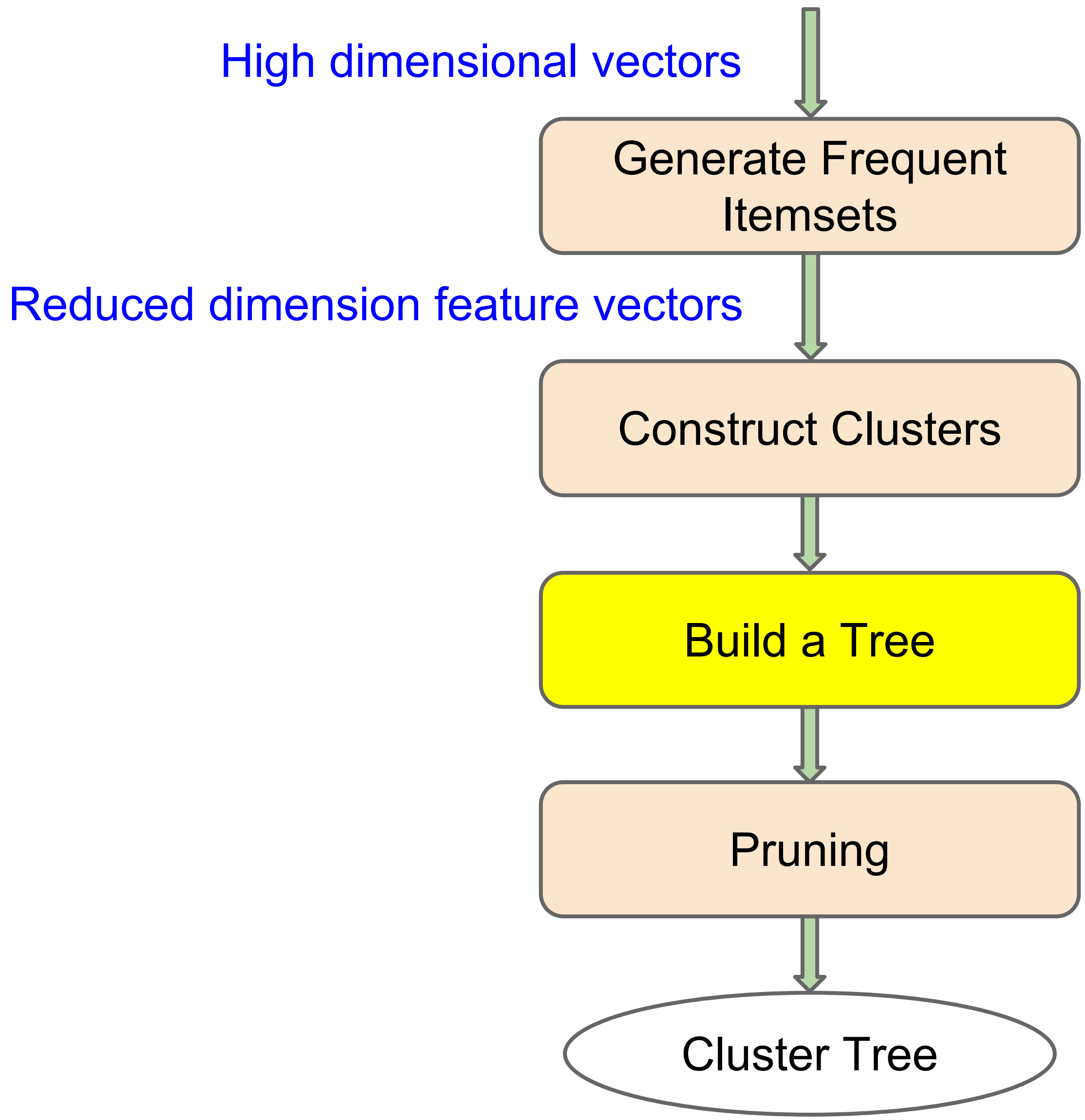


Similarity between a cluster and a document

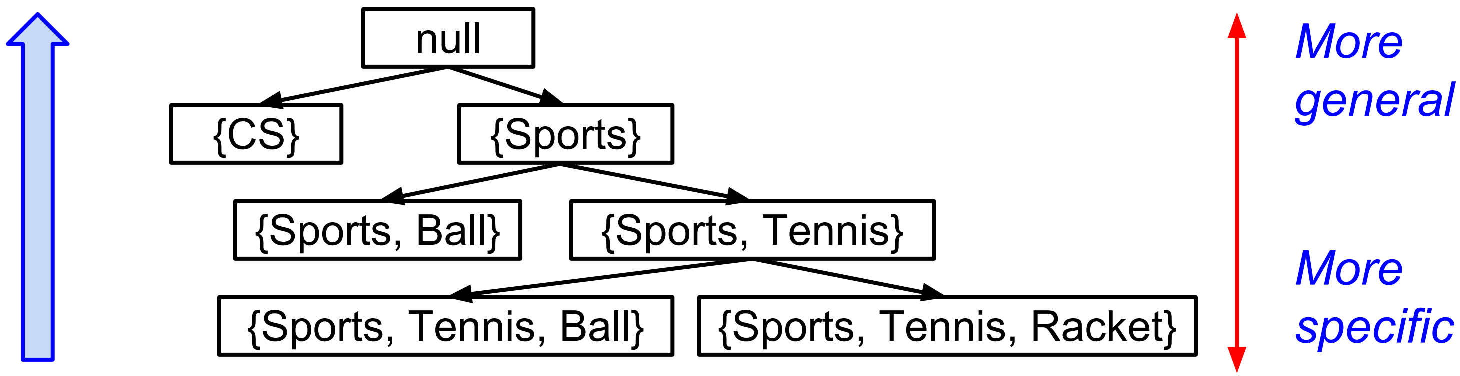
$$Score(C_i \leftarrow doc_j) = [\sum_x n(x) * cluster\_support(x)] - [\sum_{x'} n(x') * global\_support(x')]$$



FIHC Algorithm Overview

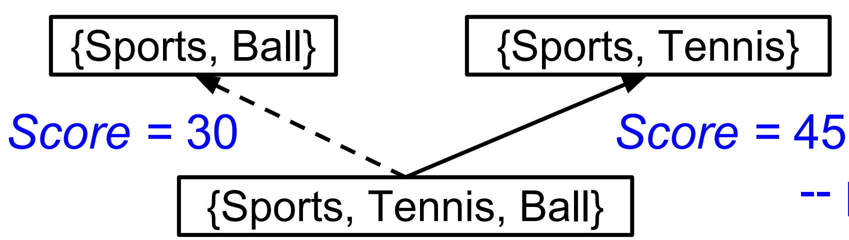


- FIHC build the hierarchical tree after clustering



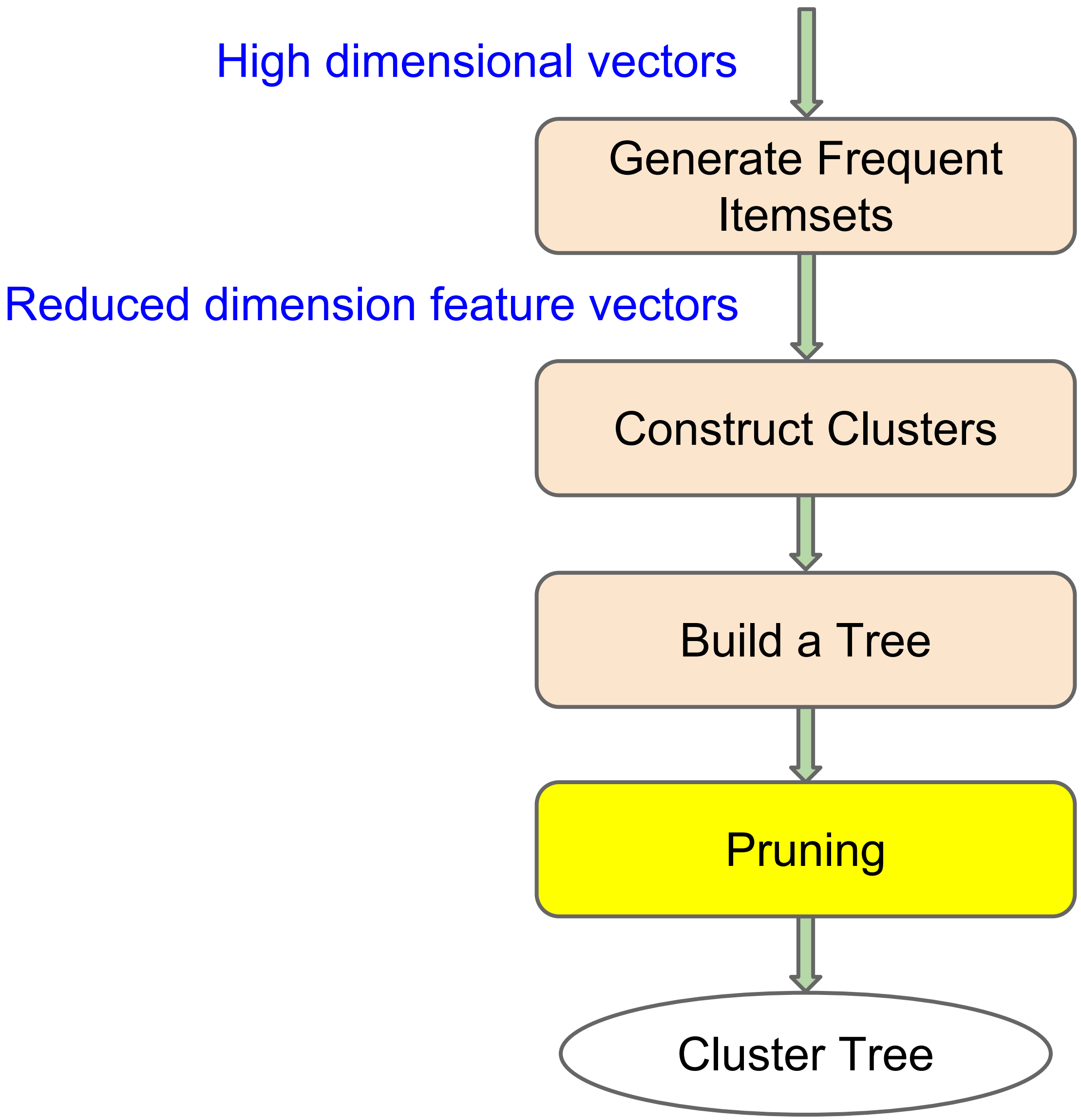
Bottom-up: Start from the largest cluster label

- How to choose the best parent?



-- pick the most similar parent

FIHC Algorithm Overview



FIHC - Cluster Tree Pruning

- Cluster Tree Pruning - Why?
  - Remove overly specific child clusters
  - Document of the same topic may distributed over different subtrees, which would lead to poor clustering quality

- Cluster Tree Pruning: Inter-Cluster Similarity

  Geometric mean of both direction

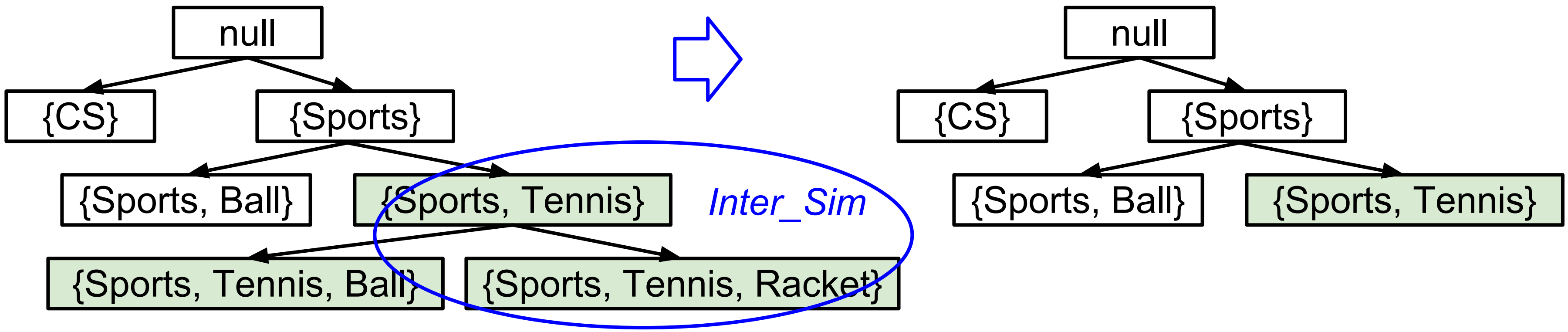
  - $Inter\_Sim(C_i, C_j) = (Sim(C_i, C_j) * Sim(C_j, C_i))^{½}$

  Treat *Cj* as a doc, reuse the score function

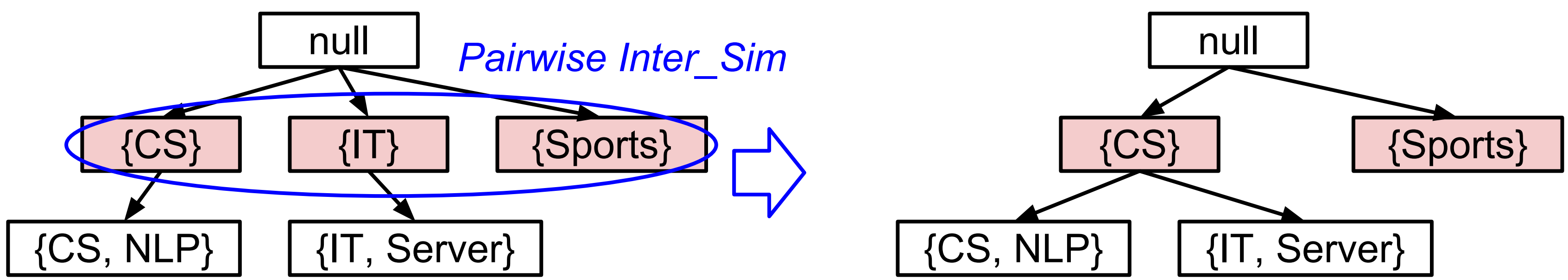
  - $Sim(C_i, C_j) = Score(C_i, doc(C_j)) / \underline{(\Sigma n(x) + \Sigma n(x'))} + 1$

  Normalized by item frequency
  x: global frequent items in both *Ci* and *Cj*
  x': global frequent items in *Cj* but not in *Ci*
  n(): frequency in *Cj*

- Cluster Tree Pruning - Child Pruning (for level > 1)
  - Shorten the tree



- Cluster Tree Pruning - Sibling Merging (for level = 1)
  - Narrow the tree



Experimental Results - Data Sets

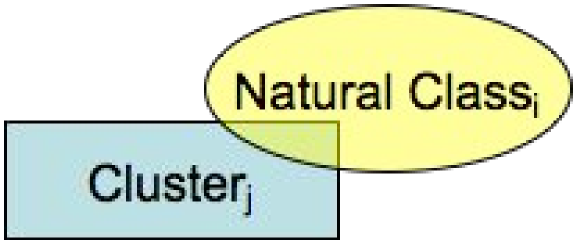
- Each document is pre-classified into a single natural class.

Data Set	Number of Documents	Number of Classes	Class Size	Average Class Size	Number of Terms
Classic4	7094	4	1033 – 3203	1774	12009
Hitech	2301	6	116 – 603	384	13170
Re0	1504	13	11 – 608	116	2886
Reuters	8649	65	1 – 3725	131	16641
Wap	1560	20	5 – 341	78	8460

Table 5.1: Summary descriptions of data sets

Clustering Quality: F-measure

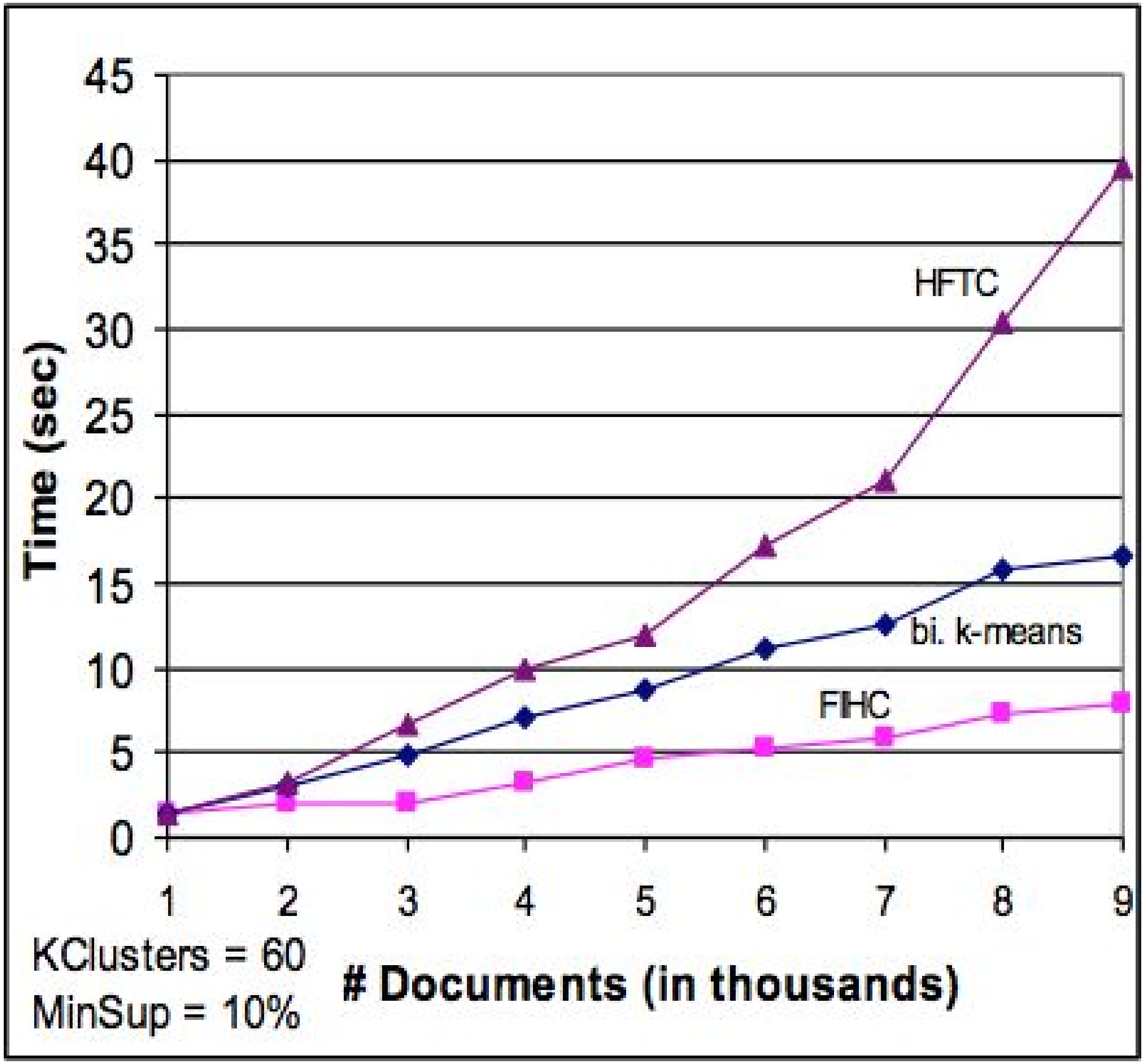
- Widely used method for evaluation
- F-measure range is from 0 to 1
- Weighted average of recall and precision



Data Set (# of natural classes)	# of Clusters	F-measure			
		FIHC	UPGMA	Bi. k-means	HFTC
Classic4 (4)	3	0.62*	×	0.59	n/a
	15	0.52*	×	0.46	n/a
	30	0.52*	×	0.43	n/a
	60	0.51*	×	0.27	n/a
	Average	0.54	×	0.44	0.61*
Hitech (6)	3	0.45	0.33	0.54*	n/a
	15	0.42	0.33	0.44*	n/a
	30	0.41	0.47*	0.29	n/a
	60	0.41*	0.40	0.21	n/a
	Average	0.42*	0.38	0.37	0.37
Re0 (13)	3	0.53*	0.36	0.34	n/a
	15	0.45	0.47*	0.38	n/a
	30	0.43*	0.42	0.38	n/a
	60	0.38*	0.34	0.28	n/a
	Average	0.45*	0.40	0.34	0.43
Reuters (65)	3	0.58*	×	0.48	n/a
	15	0.61*	×	0.42	n/a
	30	0.61*	×	0.35	n/a
	60	0.60*	×	0.30	n/a
	Average	0.60*	×	0.39	0.49
Wap (20)	3	0.40*	0.39	0.40*	n/a
	15	0.56	0.49	0.57*	n/a
	30	0.57	0.58*	0.44	n/a
	60	0.55	0.59*	0.37	n/a
	Average	0.52*	0.51	0.45	0.35

Table 5.2: F-measure comparison
× = not scalable to run * = best competitor

Experimental Results - Efficiency



Conclusions

- Main contributions of the paper
  - Using frequent itemsets to **reduce dimension**, so as to achieve higher efficiency and scalability
  - Measuring **cluster similarity** based on frequent itemsets
  - High clustering quality
  - Number of clusters is optional as input parameter
  - Meaningful cluster labels

Q & A

Thanks!