Patent Mining: Use of Data/Text Mining for Supporting Patent Retrieval and Analysis

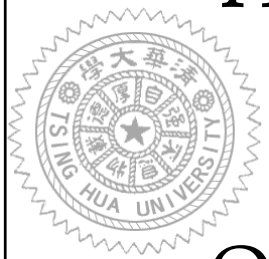
by Chih-Ping Wei (魏志平), PhD

Institute of Service Science and
Institute of Technology Management
National Tsing Hua University

December 5, 2008

國立清華大學
CTM 科技管理學院
College of Technology Management

Agenda

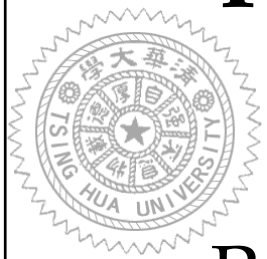


- Overview
- Study 1: Patent Prior Art Retrieval: A Text Summarization Approach
- Study 2: Patent Valuability Prediction

Patents

- A patent is a set of exclusive rights granted by a state to an inventor or his assignee for a fixed period of time in exchange for a disclosure of an invention.
- Typically, a patent application must include one or more claims defining the invention which must be new, inventive, and useful or industrially applicable.
- The exclusive right granted to a patentee in most countries is the right to prevent or exclude others from making, using, selling, offering to sell or importing the invention.

Patent Retrieval

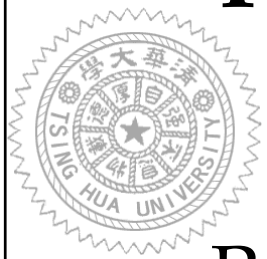


- Patent retrieval is a difficult task in the information retrieval (IR) domain: one aspect of the difficulty arises from patent document characteristics such as special stylistic features and idiosyncratic terminology.
- Patent retrieval can be classified into two major categories:
  - Technology survey search is to find all patent documents related to a subject topic or a specific technology.
  - Prior art retrieval is to identify all relevant prior arts for a given patent or patent application.

Patent Analysis

- Analysis of patents can support (Ashton and Sen, 1988; Porter and Newman, 2004):
  - R&D management
  - Mergers and acquisitions
  - Product area surveillance
  - Technology competition analysis
  - Intellectual asset management
  - Patent portfolio management
  - New venture evaluation
  - Technology foresight
  - Strategic planning, etc.

Focus of This Talk



- Prior art retrieval supported by a text mining (specifically, text summarization) approach
- Data-mining-based patent valuability prediction that could be essential to technology competition analysis, intellectual asset management, patent portfolio management, and new venture evaluation

Study 1: Patent Prior Art Retrieval: A Text Summarization Approach

Outline

- Introduction
- Design of Summary-based Prior Art Retrieval Technique
  - Overall Process
  - Patent Summarization System
  - Summary-based Prior Art Retrieval System
- Empirical Evaluation
- Conclusion & Future Research Directions

Introduction (1/3)

- A patent is a collection of exclusive rights which protect an inventor's new machine, process, article, or any new improvement theory for a fixed period of time.
- In addition to its abstract, claims, and description, a patent (or patent application) needs to include a list of cited patents (i.e., prior arts).
- Citations serve to show how the claimed invention differs from the "prior arts." The basic purpose of citing "prior arts" in a patent is to inform the public in general that such patents are in existence and should be considered when evaluating the validity of the patent claims (USPTO, 2001).

Introduction (2/3)

- Patent applicants and examiners are cautious with patent citations, which help define a patent and have direct influence on its economic value (Lai and Wu, 2005).
- Prior art retrieval refers to the process of identifying all relevant prior arts for a given patent or patent application.
- Prior art retrieval is applicable to patentability, novelty, validity, and infringement investigations when adapted to different search environments (Fujita, 2007).

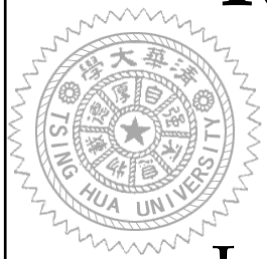
Introduction (3/3)

- Prior art retrieval requires more rigid standards of relevance (i.e., adequacy as evidential material) than a technology survey search (i.e., a subject topic search of patents). Accordingly, this leads to a smaller number of relevant prior arts for each query (Fujita, 2007).

Research Motivation and Objectives (1/2)

- Existing prior art retrieval research generally employs information retrieval (IR) techniques.
- Recent studies extend IR techniques by including additional mechanisms to improve retrieval effectiveness.
- Such mechanisms include query expansion (expanding claims with relevant sentences in the description of the query) (Konishi, 2005), citation analysis (combining the text-based similarity and the PageRank score to re-rank relevant patents) (Fujii, 2007), and cluster-based retrieval (considering the IPC classification codes of prior arts and the query) (Kang et al., 2007).

Research Motivation and Objectives (2/2)

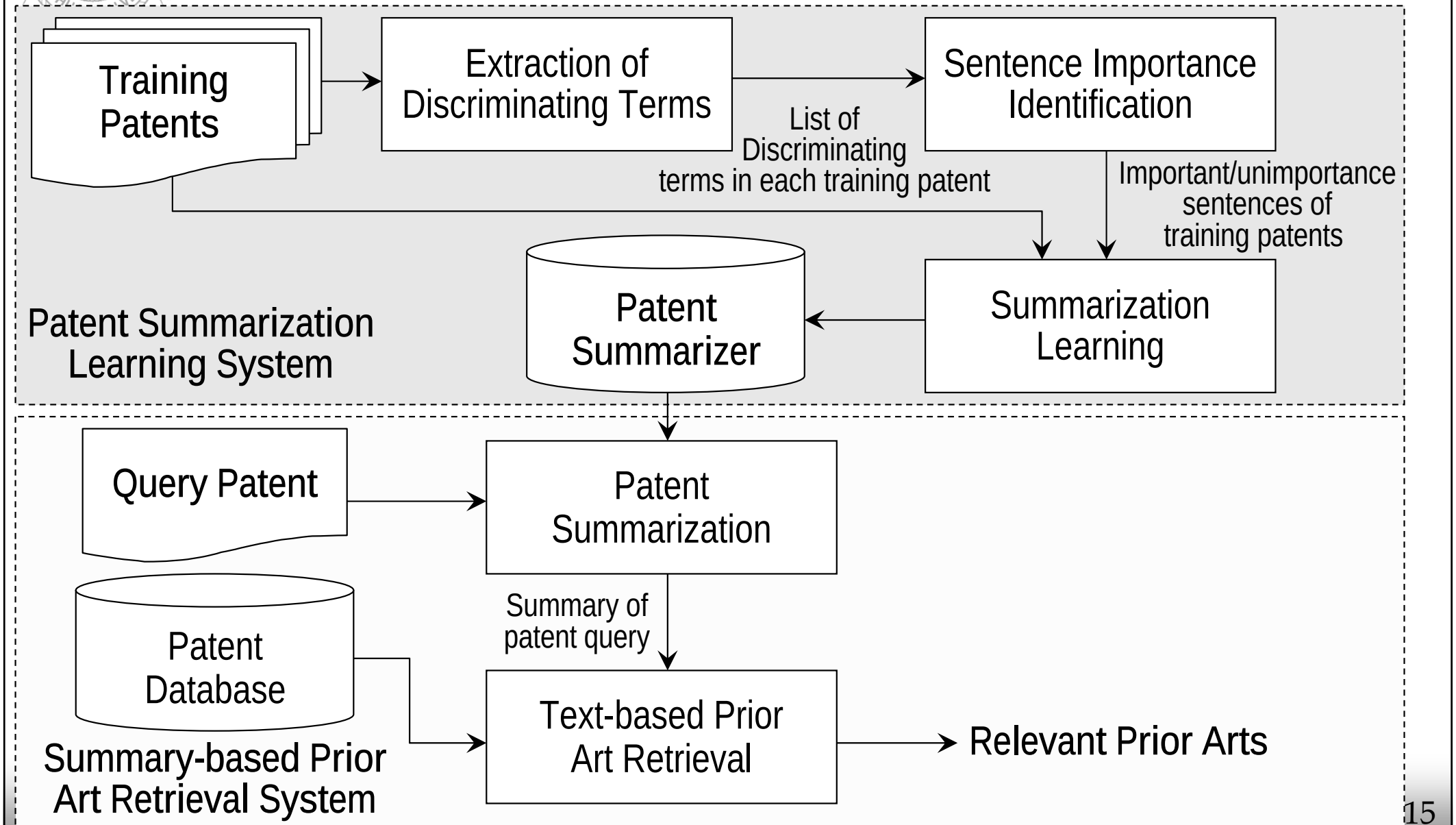


- In this study, we attempt to use the text summarization approach to support prior art retrieval.
- Our rationale is that sentences in a patent document are not equally important in describing the invention claimed in the patent.
- Thus, we employ the text summarization approach to develop an automatic patent summarization technique for summarizing important sentences in patent documents and use these patent summaries for prior art retrieval.
- The proposed technique is referred to as a summary-based prior art retrieval technique.

Research Questions

- How can we train a patent summarizer without having to manually identify important and unimportant sentences in a training patent document?
- Is the proposed summary-based prior art retrieval technique more effective than a traditional text-based prior art retrieval technique?

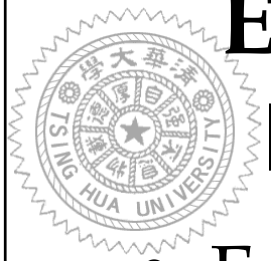
Overall Process of Summary-based Prior Art Retrieval Technique



Patent Summarization Learning System: Extraction of Discriminating Terms (1/3)

- This step is the core to automatically train a patent summarizer without having to manually identify important and unimportant sentences in a training patent.
- For each training patent (p_i), the set PA_i of prior arts cited by the patent is considered relevant to the training patent.
- We select the same number of existing patents that are highly similar (in their content) to p_i but not in PA_i and consider them (denoted as NPA_i) as the training set of non prior arts for p_i.

Patent Summarization Learning System: Extraction of Discriminating Terms (2/3)



- For each term f_{ij} (noun or noun phrase in this study) appearing in p_i, we calculate the χ^2 statistic of the term with respect to PA_i and NPA_i.

	PA_i	NPA_i
Occurrence of f_{ij}	A	B
Nonoccurrence of f_{ij}	C	D

$$\chi^2(t_{ij}, P_i) = \frac{(A+B+C+D)\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

Patent Summarization Learning System: Extraction of Discriminating Terms (3/3)

- For each training patent, we then retain all terms whose χ^2 statistic is significant at the 95% significance level.
- We consider a discriminating term as a positive cue, if ($A \times D > B \times C$) in its corresponding matrix.

Patent Summarization Learning System: Sentence Importance Identification

- For each training patent, we determine the importance of every sentence in the patent document as follows:
  - Important: if the number of positive cues in the sentence ≥ 2, the sentence is regarded as an importance one.
  - Unimportant: if the number of positive cues in the sentence = 0, the sentence is regarded as an unimportance one.
  - Unknown: otherwise.

Patent Summarization Learning System: Summarization Learning (1/4)

- After the sentence importance identification phase, we automatically obtain a set of important sentences and a set of unimportant sentences for each training patent.
- These important and unimportant sentences across all training patents become the input to the summarization learning phase.
- As with previous text summarization research, we consider four types of variables possibly affecting the importance of a sentence in a document: Location, Sentence Length, Title Word, and Thematic Word.

Patent Summarization Learning System: Summarization Learning (2/4)

Variable	Domain	Description
Location	{1, 2, 3}	The sentence occurs in the abstract, claim, or description.
Location (Abstract)	{0, 1, 2, 3}	The sentence occurs in the first 1/3 (coded as 1), first 2/3 (coded as 2), or last 1/3 (coded as 3) of sentences in the abstract section. If the sentence is not in the abstract section, it is coded as 0.
Location (Claim)	{0, 1, 2, 3}	The sentence occurs in the first 1/3, first 2/3, or last 1/3 of sentences in the claim section.
Location (Description)	{0, 1, 2, 3}	The sentence occurs in the first 1/3, first 2/3, or last 1/3 of sentences in the description section.

Patent Summarization Learning System: Summarization Learning (3/4)

Variable	Domain	Description
Sentence Length	Integer	Number of words in the sentence.
Title Word	Binary	Whether the sentence contains any title words (i.e., nouns or noun phrases) (1: yes, 0: no).
Thematic Word	Integer	the number of thematic words appearing in the sentence, where the set of thematic words are defined as the 50 most representative nouns or noun phrases measured by Okapi BM25 TF×IDF metric).

Patent Summarization Learning System: Summarization Learning (4/4)

- After we represent all important and unimportant sentences of the training patents with the above-mentioned variables, we employ C4.5 (a decision tree induction technique) and Naïve Bayes classifier as our alternative learning algorithms. The resultant model is referred to as a patent summarizer.
- Given an input patent document, each sentence is first represented with the above-mentioned variables. Subsequently, the patent summarizer will generate an importance probability for each sentence in the patent document.

Summary-based Prior Art Retrieval System: Patent Summarization

- Given a query patent, the patent summarizer generates an importance probability for every sentence in the patent document.
- On the basis of the prespecified compression ratio (*cp-ratio*) (i.e., the percentage of sentences to be retained in its summary), the top *cp-ratio* of sentences are selected and become the summary of the query patent.

Summary-based Prior Art Retrieval System: Text-based Prior Art Retrieval (1/3)

- As with all text-based prior art retrieval or general IR techniques, we perform feature selection and document representation for the target query and documents to be retrieved.
- In this study, we employ the Okapi BM25 TF×IDF metric commonly adopted for feature selection.

$$TF \times IDF_{BM25}(f,d) = \frac{(k1+1)\,freq(f,d)}{k1\left((1-b)+b\dfrac{dl(d)}{avdl}\right)+freq(f,d)} \times \left(k4+\log\frac{N}{df(f)}\right)$$

Summary-based Prior Art Retrieval System: Text-based Prior Art Retrieval (2/3)

Notations in Okapi BM25 TF×IDF metric:

f	a term in the target patent
d	the target patent
N	the total number of patents in the collection
df(f)	the number of patents in which *f* appears
freq(f,d)	the number of occurrences of *f* in *d*
dl(d)	the document length of *d*
avdl	the average document length in the collection
*b*1, *b*, *k*4	parameters. In this study, we set *k*1 as 1.2, *b* as 0.75, and *k*4 as 0, as suggested by prior studies.

Summary-based Prior Art Retrieval System: Text-based Prior Art Retrieval (3/3)

- For the summary of the query patent and for each existing patent in the database, we select the top *k* features with the highest Okapi BM25 TF×IDF metric scores to represent the patent.
- The Okapi BM25 TF×IDF metric is also adopted as the document representation scheme.
- Subsequently, we employ the cosine measure to estimate the similarity between the query patent and any existing patent in the database.
- Finally, we retrieve the top-*n* patents as the relevant prior arts for the query patent.

Empirical Evaluation: Data Collection (1/2)

- Our dataset includes 68,993 USPTO patents from 1990 to 2006 that belong to the CCL (US Classification) class 438 (Semiconductor Device Manufacturing: Process) or its subclass.
- We randomly select 150 patents from our patent database that were issued in 2004 and had 10 to 50 prior arts as our training patents.
- We use 300 patents issued in 2005 as the query patents for evaluating the effectiveness of the proposed summary-based prior art retrieval technique.

Empirical Evaluation: Data Collection (2/2)

- In addition, for each training patent or query patent, we add its prior arts that do not belong to CCL class 438 into our patent database.
- As a result, additional 9,232 patents are included in our dataset.

Empirical Evaluation: Performance Benchmark

- We use the traditional full-text-based prior art retrieval technique as our benchmark technique.
- Specifically, instead of using summaries of query patents, our benchmark technique uses
  - the full text of each query patent as the input for prior art retrieval
  - Okapi BM25 TF×IDF metric for feature selection and document representation
  - cosine measure to estimate the similarity between a query patent and any existing patent.

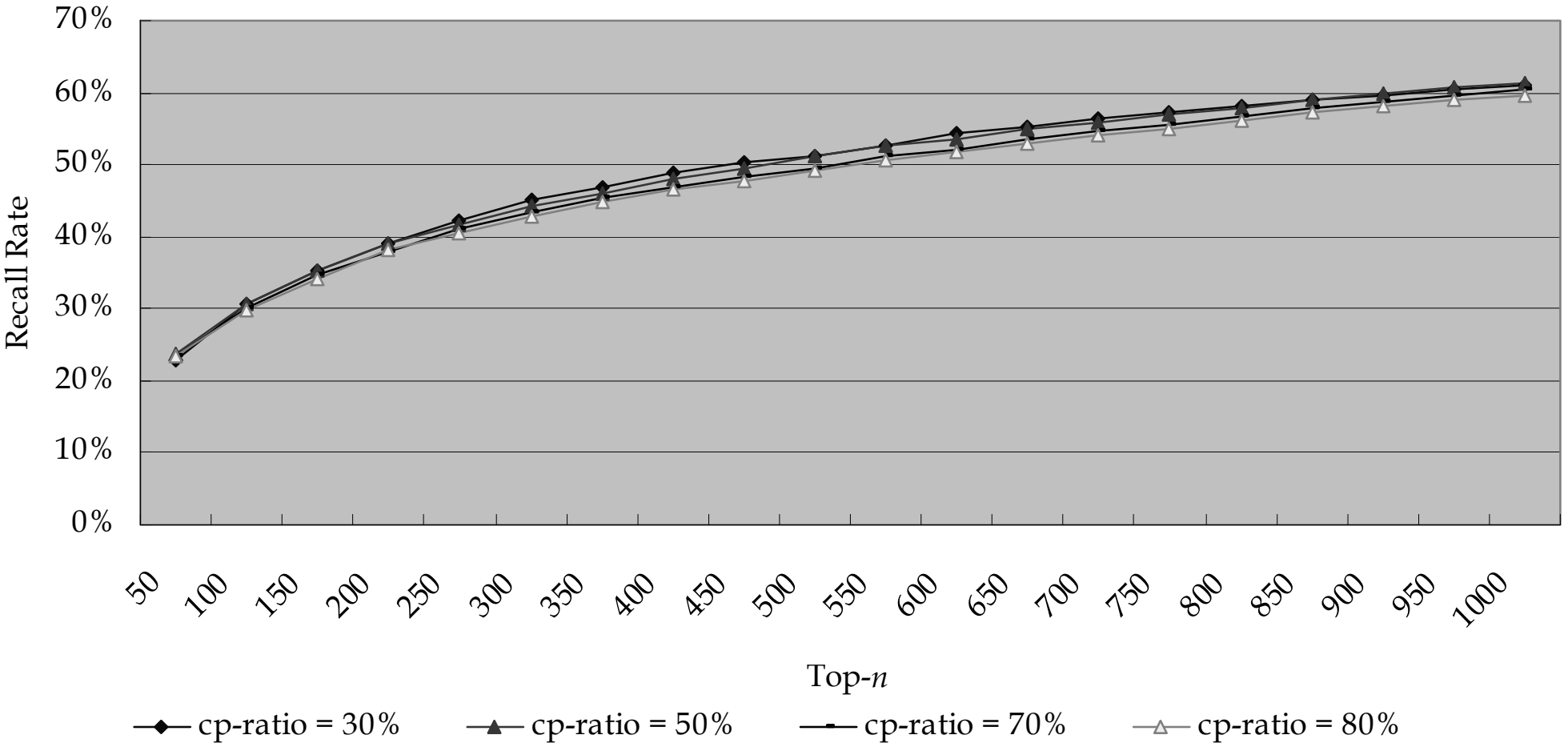
Empirical Evaluation: Evaluation Criteria

- We use the top-*n* recall rate to measure the effectiveness of each technique under investigation.

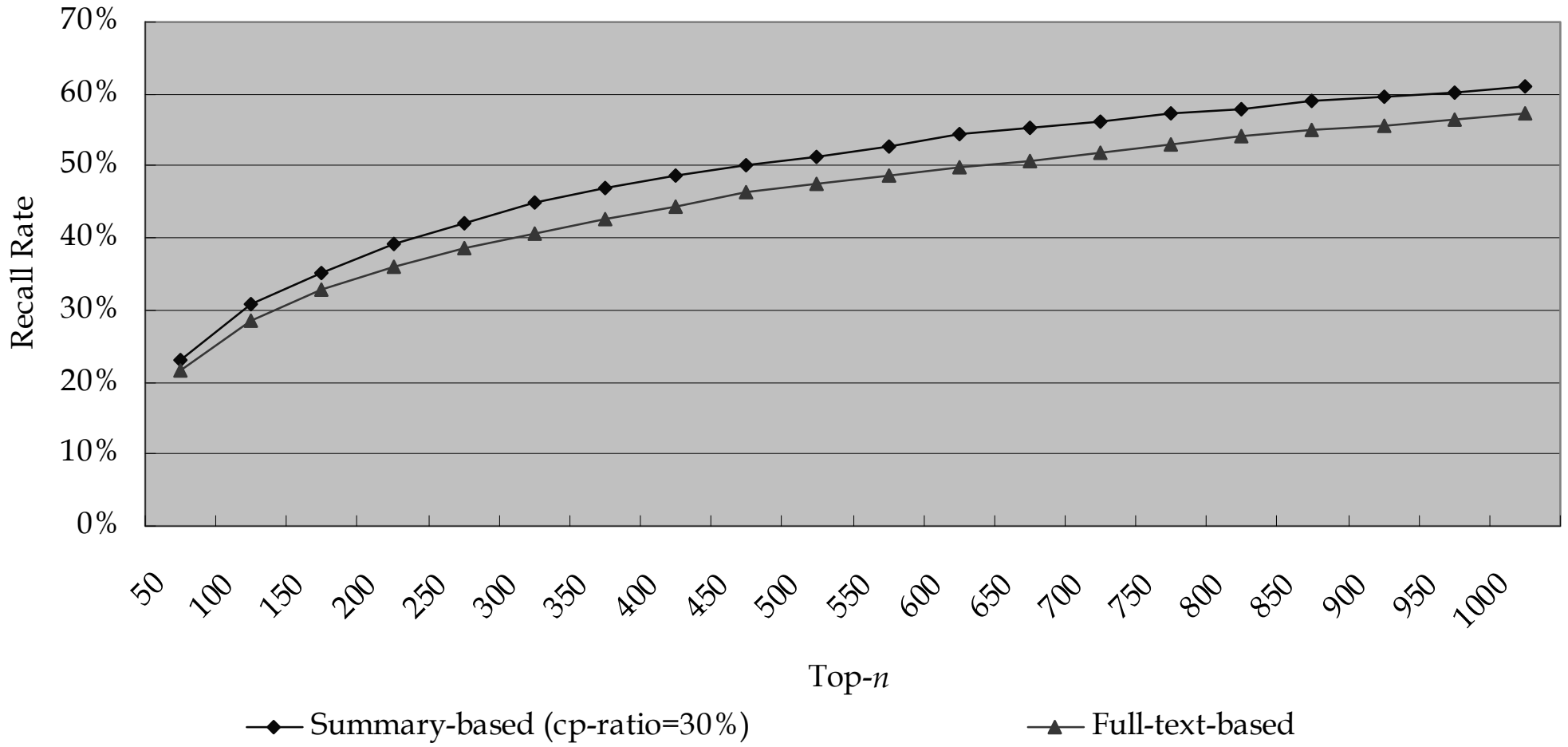
Empirical Evaluation: Evaluation Results

- Our preliminary evaluation results suggest that the use of Naïve Bayes classifier for summarization learning outperforms that of C4.5.
- Thus, we use Naïve Bayes classifier for summarization learning in subsequent experiments.

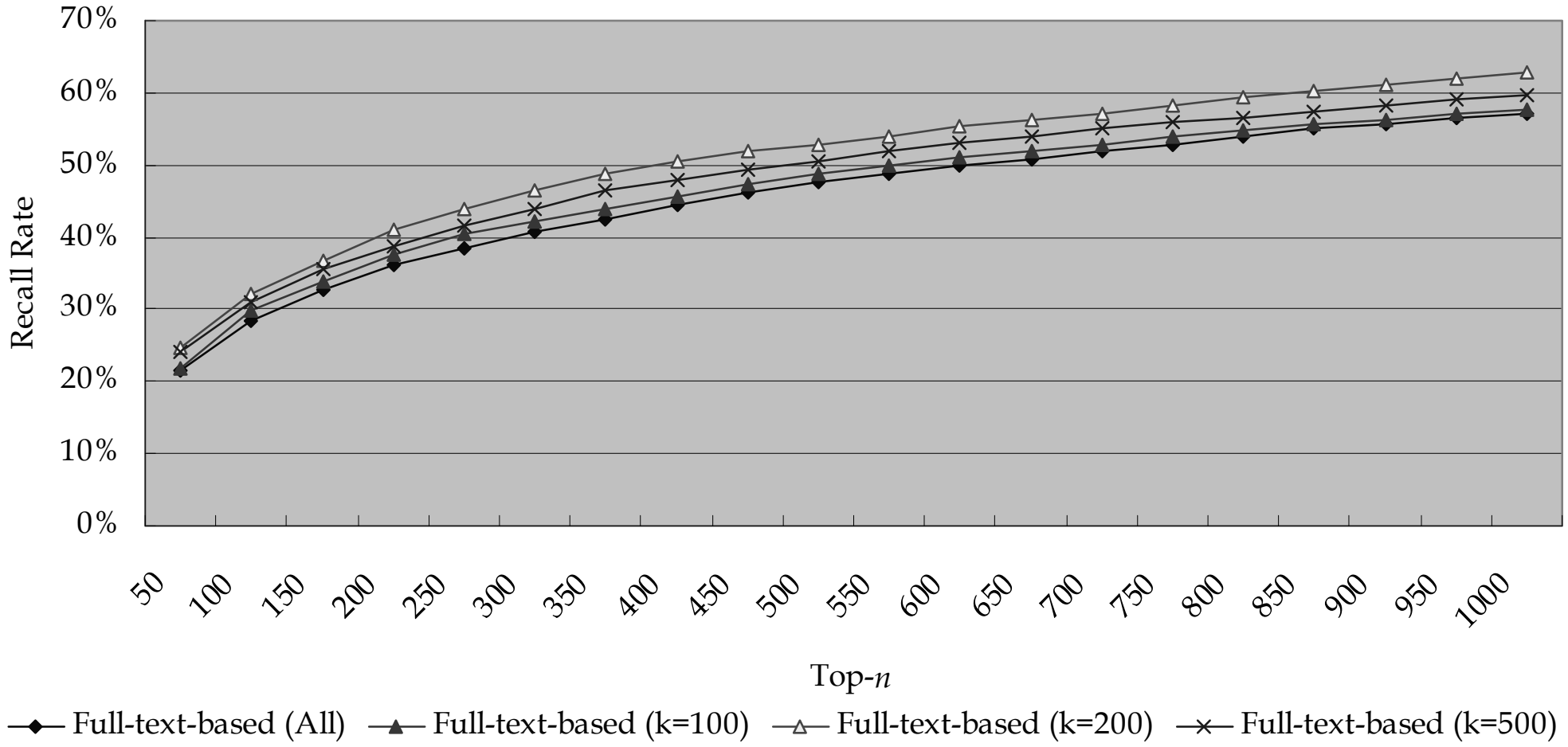
Effects of Compression Ratio (Without Feature Selection)



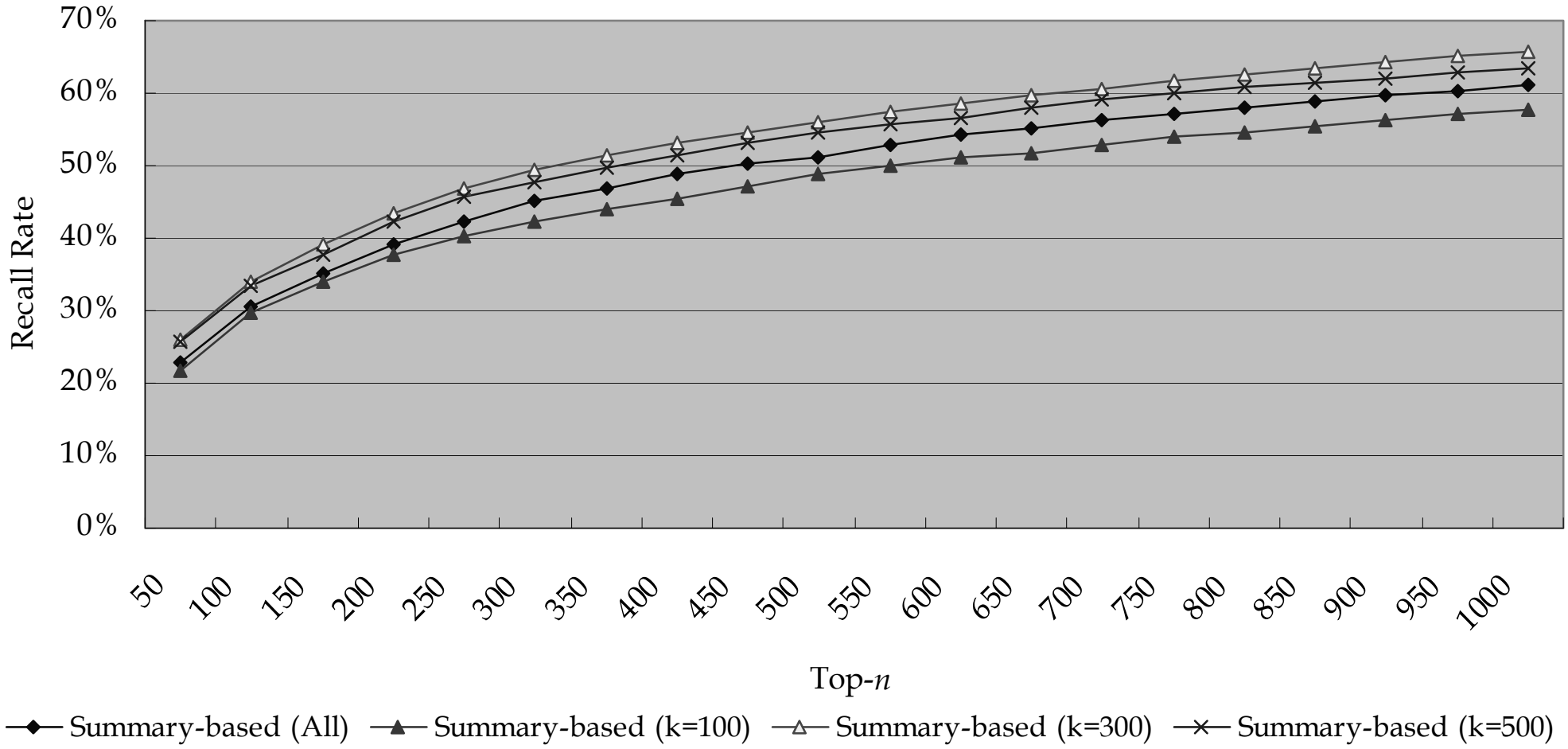
Comparative Evaluation (Without Feature Selection)



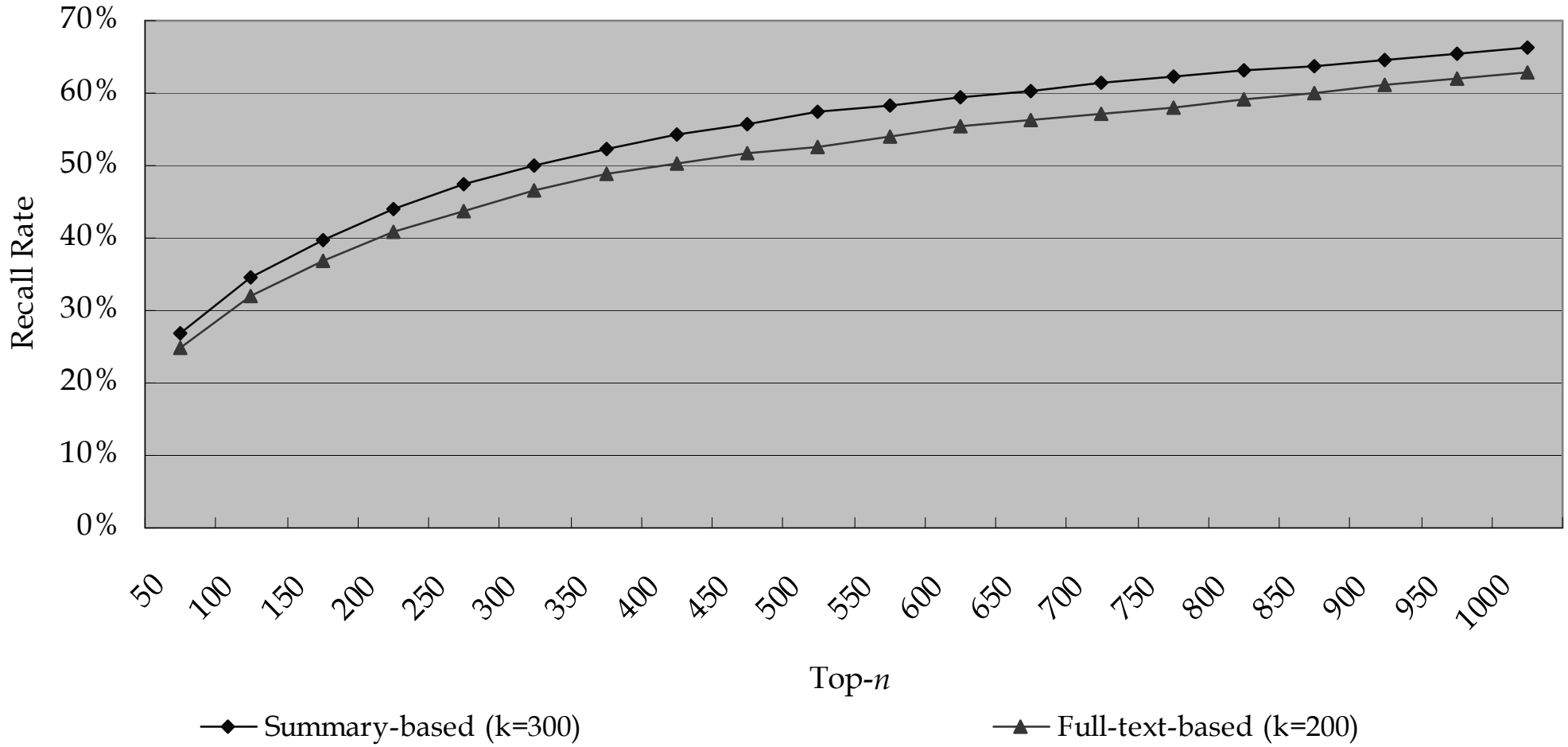
Effects of Feature Size on Benchmark Technique



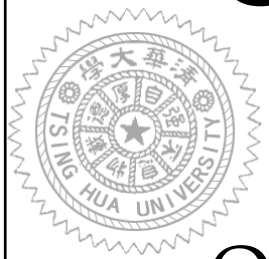
Effects of Feature Size on Summary-based Prior Art Retrieval Technique (*cp-ratio*=0.3)



Comparative Evaluation (With "Optimal" Feature Size)



Conclusion and Future Research Directions



- Our empirical evaluation results suggest that the proposed summary-based prior art retrieval technique outperforms the traditional full-text-based technique.
- Future research directions include:
  - Extending the proposed technique with the consideration of IPC class information and with the query expansion approach
  - Extending the current empirical evaluations with patents in other classes

Study 2: Patent Valuability Prediction

Outline

- Introduction
- Preliminary Set of Variables and Learning Algorithm
- Preliminary Evaluations
- Future Research Directions

Introduction (1/4)

- Some patents are intrinsically more valuable than others. Many patents are not worth enforcing—either because the inventions they cover turn out to be worthless or because even if the invention has economic value the patent does not.
- As a result, companies or inventors are facing a challenging issue: of the patents they have (or are interested), what are valuable and what are not?
- This study focuses on applying a data mining technique for patent valuability prediction, defined as "classifying a patent into one of predefined value categories (e.g., valuable and non-valuable)."

Introduction (2/4)

- Patent valuability prediction can support, fully or partially, the following decision scenarios:
  - Whether to continue maintaining a patent (by patent assignees)?
  - Which patents are good candidates for licensing or sale (by patent assignees, especially universities or research institutes)?
  - Whether a technology-based new venture is worth investing (by investors or venture capital)?

Introduction (3/4)

- Prior studies focus on suggesting proxies for valuable and non-valuable patents. Such proxies include litigated patents (for valuable) vs. non-litigated patents (for non-valuable) or continuing-to-maintain patents (for valuable) vs. failed-to-maintain patents (for non-valuable).
- These studies also analyze the importance of variables in classifying the two value categories.
- Because the variables are developed with respect to their proxies, the set of variables tend to be not comprehensive and some of them (e.g., owner status) may not useful for patent valuability prediction.

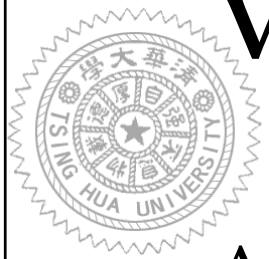
Introduction (4/4)

- Moreover, prior studies conduct their analyses without using patents' "true" value categories (i.e., solely relying on proxy data). This greatly limits the external generalizability of these studies.
- This study attempts to fill up the aforementioned gaps of prior studies by:
  - Using a more comprehensive list of variables for predicting patent value category
  - Employing a data mining technique to construct a patent valuability prediction model
  - Evaluating the effectiveness of the induced model using patents' "true" value categories

Preliminary Set of Variables for Patent Valuability Prediction (1/2)

- Number of inventors
- Number of claims
- Number of assignees
- Number of foreign patents
- Number of backward citations
- Number of backward US citations
- Number of backward foreign citations
- Number of backward non-patent citations

Preliminary Set of Variables for Patent Valuability Prediction (2/2)



- Average number of forward citations per year
- …
- Number of primary classes
- Number of classes of forward citations
- …
- Years between file and issue date
- Years to expire
- …

Learning Algorithm (1/3)

- We employ C4.5 (a decision-tree induction technique) as our underlying learning algorithm.
- Specifically, C4.5 uses a set of training examples preclassified into “valuable” or “non-valuable” category to induce a classifier in a decision tree structure.
- In the prediction process, the classifier produces the probability of being valuable for a given input patent.
- Because it is easier to obtain non-valuable patents than valuable ones, the categories of a set of training examples tend to be asymmetric or highly skewed.

Learning Algorithm (2/3)

- To deal with such skewness of categories in a training set, we develop an ensemble approach for learning.
- Assume that the valuable category is a minority class and the number of valuable patents for training is n.
- The n valuable patents are included in a training subset. Given a prespecified ratio (e.g., 1: α), we randomly sample $n \times \alpha$ "non-valuable" patents and also include them in the training subset.
- Subsequently, we employ C4.5 to induce a (base) classifier from this training subset.

Learning Algorithm (3/3)

- We repeat this training-subset-generation and base-classifier-induction process *k* times, thus producing *k* base classifiers. (In this study, we set *k* as 30).
- During the prediction process for an input patent, the *k* classifiers make predictions independently. The overall prediction (i.e., probability of being valuable) of the input patent is the average of the probabilities produced by the *k* base classifiers.

Preliminary Evaluations: Collection of Patents for Training Purposes

- We use proxies for collecting patents (referred as the proxy dataset) and employ them for training purposes.
- Specifically, litigated patents are collected and considered as valuable patents. In contrast, patents failed to maintain by firms of the litigated patents previously collected are regarded as non-valuable patents.
- Accordingly, we collect 37 valuable patents and 1,249 non-valuable patents from 14 companies (e.g., TSMC, UMC, IBM, HP, etc).

Preliminary Evaluations: Collection of Patents with "True" Value Category

- We use the expert judgment approach to obtain the "true" value category for 31 US patents assigned to a national university in Taiwan.
- For each patent, a domain expert evaluates the patent and assigns a score between 0 and 100 along 11 dimensions (e.g., technological innovation, commercialization possibility, protection scope, etc).
- The overall score is then the weighted average of the scores on these 11 dimensions.
- Finally, on the basis of the overall scores, the 31 US patents are classified into three categories: valuable (5), moderate (16), and non-valuable (10).

Preliminary Evaluation Results: Cross Validation on Proxy Dataset

Ratio in Training Subset	Valuable Category		Non-valuable Category		Overall Accuracy
	Precision	Recall	Precision	Recall	
1:1	69.28%	69.28%	69.45%	69.28%	69.28%
1:2	56.69%	51.53%	76.86%	80.23%	70.66%
1:3	51.63%	45.41%	82.48%	85.62%	75.56%

- The evaluation results are encouraging (i.e., achieving 69.28% to 75.56% in overall accuracy).
- Because the training subset using 1:1 ratio attains the highest precision and recall for the valuable category, we adopt this ratio for subsequent experiments.

Preliminary Evaluation Results: Evaluation Against Expert Judgments

Contingency Table

		Predicted Value Category		
		Valuable	**Moderate**	**Non-valuable**
Actual Value Category	**Valuable**	3	2	0
	Moderate	2	10	4
	Non-valuable	2	7	1

- We use 0.33 and 0.66 as probability cutoff points when assigning patents to the three value categories.
- The overall accuracy is 14/31 = 45.16%.
- Only two predictions are considered "significantly off" (i.e., predicting 2 non-valuable patents as valuable).

Future Research Directions

- Continue to improve the effectiveness of the proposed patent valuability prediction technique by incorporating additional variables.
- Perform a larger scale of evaluation that involves more number of patents with “true” value category.
- Extend the proposed valuability prediction technique to patent applications or even patent proposals (for internal reviews by companies).