Data & Knowledge Engineering xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

An effective High Recall Retrieval method

Justin JongSu Song, Wookey Lee*, Jafar Afshar

Department of Industrial Engineering, Inha University, Incheon, South Korea

ARTICLE INFO

Keywords:
High Recall Retrieval problem
Patent retrieval
Dynamic retrieval
Independent Dominating Set Problem

ABSTRACT

The High Recall Retrieval (HRR) problem is one of the fundamental tasks for many applications such as patent retrieval, legal search, medical search, marketing research, charging and collecting tax, and literature review, etc. Given the data set obtained by the user's query, the HRR problem is defined as finding the full set of relevant documents while less review effort will be required. It is very expensive to review a lot of documents since most of the reviewers are experts in the specific fields such as patent attorneys, lawyers, marketing, and medical professionals. However, the existing HRR methods have been far from satisfactory to make them enumerate all relevant documents. This is due to the fact that not only the sheer volume of documents inevitably including noises (non-relevant documents) but also the threshold measurements have been inadequately adopted. To deal with these problems, we propose a novel solution to efficiently find all the relevant documents among a large set of results. It consists of two steps: (a) to effectively classify the entire documents and (b) to select the representative documents in each class. We formalized the problem and theoretically verified the upper-bound of our method. In the experiments, our method is more efficient than the state-of-the-art query expansion methods.

1. Introduction

High *precision* in traditional information retrieval systems has been important to find the most relevant targets, and it may be acceptable even if some best ones will be omitted. *Recall* in Patent IR (PaIR), however, is a critical issue, since the prior-art search should have been conducted without exceptions before making a full-scale business investment decision. Because a patent which is missed in the searching procedure and the infringement of the missing one might cause an enormous risk usually including a huge settlement cost, the production indemnification, and the amount reimbursed regarding the degraded Corporate Identity value.

The High Recall Retrieval (HRR) problem is one of the fundamental tasks for many applications such as patent retrieval, legal searches, medical searches, marketing research, charging and collecting taxes, and literature review, etc. These can be exemplified by situations such as when a patent examiner needs to identify all relevant patents; a lawyer needs to find every piece of evidence related to his/her case from documents that are under a legal hold; a scientist does not want to miss any piece of prior work related to his/her ongoing research; the National Tax Service imposes duties on all taxpayers exclusively.

The conventional information retrieval systems have not satisfied HRR problem, and their main purpose of them has long been focused on to maximize the precision. In HRR problem, to reduce the review efforts for the users and also not to miss any relevant results, the HRR experts have inevitably been conducted the query expression techniques which means that queries have consisted of many keywords that have repeated a lot of reformulation steps. This process has been required tedious efforts by the domain experts because the retrieval quality has been based on their ability to skillfully construct variations of queries and after that to laboriously

* Corresponding author.
E-mail addresses: jaegal83@inha.edu (J.J. Song), trinity@inha.edu (W. Lee), jafar.afshar@inha.edu (J. Afshar).

http://dx.doi.org/10.1016/j.datak.2017.07.006

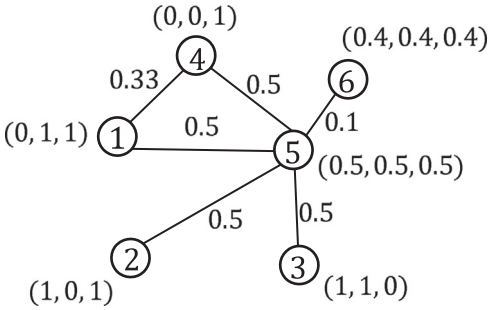


Fig. 1. Example for the limitations of existing retrieval methods.

investigate the retrieved documents. Unfortunately, the larger the collection of a target database, the more cumbersome the investigation efforts and the lower the resulting quality.

Basically, the keyword based approaches by the traditional IR methodologies have been not made successful to address the patent databases. The main reasons can be summarized as follows: (1) New keywords have been coined from each and every patent since the patent per se will be invented by a new idea or new technology. (2) The patents have been composed to hide their core keywords intentionally since they might not have been retrieved by the keyword search engine as much as possible. And (3) similar to the second reason, the patents have been tried to hide important concepts by utilizing 'common keywords', which will be shuffled with a lot of noise keywords. Thus, the traditional keyword-based IR technologies, as well as the cutting edge ones, have been far from the patent retrieval requirements.

As an example, consider Fig. 1 (nodes only, ignore the links that will be discussed in the next paragraph again) with a total of six documents. Assume that node 6 is a relevant document. The conventional Relevance Feedback approach determined by *Rocchio* method [1] will decide that node 1, 2, and 3 are individually turned out to be "not relevant." Accordingly, the IR system will try to find the farthest node (excluding the relevant documents) from the average values of nodes 1, 2, and 3 (0.67, 0.67, 0.67) so that the node 4 will be selected and it also is not relevant, so that the process will be repeated and no relevant document is found, until all the node has to be enumerated. This means that the *Rocchio* based IR system will check all documents independently and will not be able to find the right solution properly since the relevance feedback results will return an inexact threshold. Basically, this phenomenon comes from the non-relevant document decision with parameter γ in the *Rocchio* formula.

However, the method we propose is capable of finding relevant documents as quickly as possible which focuses on the central nodes (representative node) and a distance threshold. Now, consider Fig. 1 as a graph model (nodes and links together) and assume that the distance are calculated using the Edit distance. (Our method is not restricted by a certain distance metric.) Suppose that the distance threshold $\varepsilon = 0.5$, node 5 is the central node, and the Edit distance between random nodes such as $d(1, 4)$, $d(1, 5)$, $d(5, 6)$ are 0.33, 0.5, 0.1, respectively ($d(1, 4) = (|0 - 0| + |0 - 1| + |1 - 1|)/3 = 0.33$, $d(1, 5) = (|0 - 0.5| + |1 - 0.5| + |1 - 0.5|)/3 = 0.5$, $d(5, 6) = 0.1$). Based on the given threshold and the computed distances, we can find out that node 1 and 6 are not identifiable from each other since the distance between them does not satisfy the distance threshold, $d(1, 6) = 0.6 > 0.5$. Thus, the process of finding node 6 (relevant document) is performed without checking all documents independently, since it is identifiable from the central node (node 5).

Our approach is the first one for the effective examination of HRR problem with respect to 100% recall. The existing research has mostly focused on increasing recall value, but they have not been exploited to the HRR problem nor tacked to reduce the examination costs. In this study, however, we try to solve the HRR problem and to verify our method with the examination costs, that is the metric that how quickly the all relevant documents are detected.

The HRR problem with the supervised learning has been addressed to separate the relevant documents from non-relevant documents which bisect the whole document hyperplane. In supervised learning, each of the pre-selected set of documents (the "training set") is labeled as relevant and is trained by a machine-learning algorithm, which then classifies or ranks the documents in a corpus (the "test set") according to their likelihood of relevance. The fundamental limitation of the supervised learning approach [2] is that is valid only in the binary case of "relevant or not" so that multi-topics of patents can not be covered. Multiple topics and multiple categories are a prerequisite for real patents since almost all patents are relevant to multiple particular IPC (International Patent Classification) codes. Another weakness of the method is For high-recall tasks, the opinion of an expert has been provided, giving their personal decision of relevance. But, obtaining authoritative opinions for even a small set of documents may have a fluctuation, or may incur unacceptable costs and time.

To the best of our knowledge, this kind of research is unprecedented. The nearest work to our study is ReQ-ReC (ReQuery-ReClassify) [3]. This research considered a scenario where a searcher requires both high precision and high recall from an interactive retrieval process. When accessing to the entire data set, an active learning loop was used to ask for additional relevance feedback labels to refine the classifier. This model uses a representational method of relevance feedback, as Rocchio [1], and machine learning method, as Support Vector Machine (SVM). The method has a restriction in the kernel function for multi-dimensions since the HRR targets such as patent documents are definitely related to multiple classifications. The ReQ-ReC, however, is only valid for a single classification that is not realistic.

Contributions To overcome the mentioned difficulties, we propose a dynamic effective method for HRR problem where our

main contributions are summarized as follows:

- We develop an effective dynamic retrieval technique for HRR problem.
- We formalize a diverse retrieval method based on graph theoretic approach.
- We provide an efficient algorithm to remove the documents that cannot be among k relevant ones, which can minimize the reviewing time.
- The benefits of the above features are verified through conducting experiments using various datasets.

Organizations: The rest of this paper is organized as follows. An overview of High Recall Retrieval framework and its key components is given in Section 2. High Recall Retrieval with a Single-Step (HRR^1) and Double-Step (HRR^2) describe in Section 3 and 4, respectively. Section 5 simply describes the evaluation metrics for high recall retrieval. Moreover, experimental studies are given in Section 6. We discuss related work in Section 7. Finally, we conclude the paper in Section 8.

2. Problem statements

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ denote the results set obtained by a user query, and $\widetilde{\mathcal{D}} = \{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_k\}$ is the set of relevant documents, where $\tilde{d}_i \in \widetilde{\mathcal{D}}$, $\widetilde{\mathcal{D}} \subseteq \mathcal{D}$ and $1 \le k \le n$. For each $d_i \in \mathcal{D}$, the relevancy of the document is unknown, unless the document is reviewed. The relevant score of d_i is denoted as $rel(d_i)$ which is equal to 1 if d_i is the relevant document and 0 otherwise.

Problem 1. (High Recall Retrieval Problem) Given the data set $\mathcal{D}$, the high recall retrieval is represented by $S(\mathcal{D})$, then $|S(\mathcal{D})|$ numbers in needed to satisfy k relevant documents for $k \le l \le n$ ($\widetilde{\mathcal{D}} \subseteq S(\mathcal{D})$) The goal is to minimize the number of reviewed documents l for satisfying k relevant documents.

2.1. High Recall Retrieval on a graph

Suppose that $\mathcal{D}$ is a set of documents used for graph representation. For a real number ε, $\varepsilon \ge 0$, $\mathcal{N}_\varepsilon(d_i)$ as we denote a set of *neighbors* of document $d_i \in \mathcal{D}$ (i.e., the document placed in at most ε from d_i) as follows:

Definition 1 (ε*-Neighborhood*). Let $\mathcal{D}$ be a set of documents and ε, $\varepsilon \ge 0$, a real number. The ε*-Neighborhood* of a document d_i is defined by

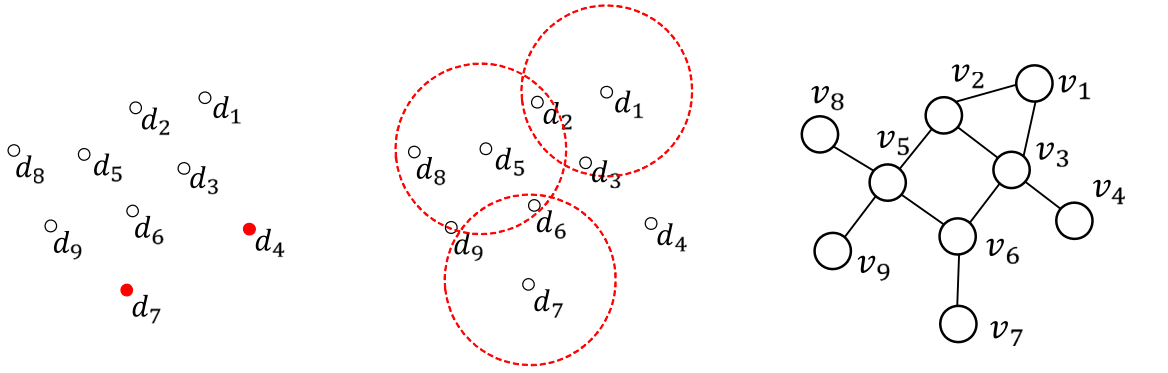
$$\mathcal{N}_\varepsilon(d_i) = \{d_j \in \mathcal{D} | dist(d_i, d_j) \le \varepsilon\} \tag{1}$$

Note that $\mathcal{N}^+_\varepsilon(d_i)$ denotes the set $\mathcal{N}_\varepsilon(d_i) \cup \{d_i\}$, i.e., the neighborhood of d_i including d_i itself. We assume that the documents in the neighborhood of d_i are considered similar to d_i, while the documents outside its neighborhood are considered dissimilar to d_i. For example, in Fig. 2(b), $\mathcal{N}_\varepsilon(d_1) = \{d_2, d_3\}$, therefore $\mathcal{N}^+_\varepsilon(d_1) = \{d_1, d_2, d_3\}$. As another example, $\mathcal{N}_\varepsilon(d_7) = \{d_6\}$, and so $\mathcal{N}^+_\varepsilon(d_7) = \{d_6, d_7\}$.

Let $G_{\mathcal{D},\varepsilon} = (V, E)$ be an undirected and weighted graph such that there is a vertex $v_i \in V$ for each document $d_i \in \mathcal{D}$ and an edge $(v_i, v_j) \in E$, if and only if, the $dist(d_i, d_j) \le \varepsilon$ for the corresponding documents d_i and d_j, $d_i \ne d_j$. Each node in G represents a document in V. For example, graph G (Fig. 2(c)) of $\mathcal{D}$ (Fig. 2(a)).

In the following, terms *document* and *node* are used interchangeably ($\mathcal{D} = V$). From now on, we use G instead of $G_{\mathcal{D},\varepsilon}$ for the rest of the paper. We desire to select a representative set, $\mathcal{S}$ of the document such that each document from $\mathcal{D}$ is represented by a similar document in $\mathcal{S}$ and the documents selected to be in $\mathcal{S}$ are dissimilar to each other. We define a Representative Sampling (*ReS*) set as follows:

Definition 2 (ε*-ReS*). Let $\mathcal{D}$ be a set of documents and ε, $\varepsilon \ge 0$, a real number. A subset $\mathcal{S} \subseteq \mathcal{D}$ is an ε-Representative Sampling set



(a) The total result set $\mathcal{D}$ (b) Making a connected graph $G_{\mathcal{D},\varepsilon}$ (c) Making a connected graph $G_{\mathcal{D},\varepsilon}$

Fig. 2. The process of making a connected graph $G_{\mathcal{D},\varepsilon}$.

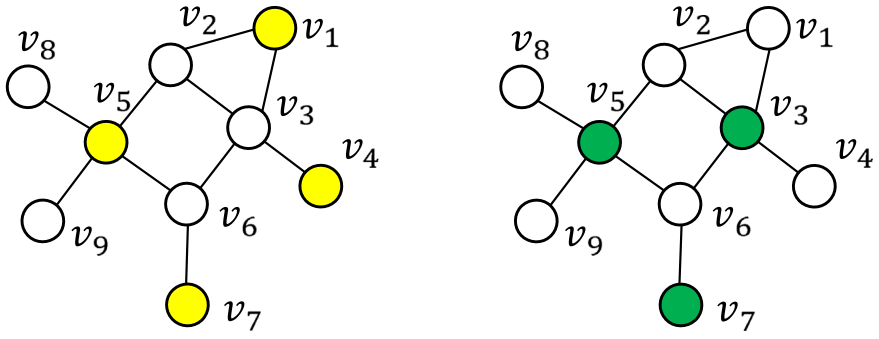


Fig. 3. *ReS* documents on G.

(ε-*ReS*) of $\mathcal{D}$, if the two following conditions hold: (i) (coverage condition) $\forall\, d_i \in \mathcal{D}$, $\exists\, d_j \in N_\varepsilon^+(d_i)$, such that $d_j \in \mathcal{S}$ and (ii) (dissimilarity condition) $\forall\, d_i, d_j \in S$ with $d_i \neq d_j$, $dist(d_i, d_j) > \varepsilon$.

The first condition ensures that all documents in $\mathcal{D}$ are represented by at least one similar document in $\mathcal{S}$ and the second condition ensures that the documents in $\mathcal{S}$ are dissimilar to each other. We call each document $d_i \in \mathcal{S}$, ε-*ReS* diverse document and ε, the *radius* of $\mathcal{S}$. When the value of ε is derived from context, we simply refer to ε-*ReS* diverse documents as diverse documents. For example, in Fig. 3(a)(b), ε-*ReS* set can be $\mathcal{S} = \{d_1, d_4, d_5, d_7\}$ or $\mathcal{S} = \{d_3, d_5, d_7\}$. In a given set $\mathcal{D}$, there may exist different cases of $\mathcal{S}$, however, we aim to select the smallest number of diverse documents and this is defined as follows.

Definition 3 (*Minimum ε-ReS Diverse set***).** A given set $\mathcal{D}$ of documents and a radius ε, an ε-*ReS* diverse set $\mathcal{S}^*$ of $\mathcal{D}$ satisfies that $|\mathcal{S}^*| \leq |\mathcal{S}|$.

Note that the number of minimum ε-*ReS* diverse set is not necessarily one and there might exist more than one in $\mathcal{D}$. For instance, the *Minimum ε-ReS Diverse set* in Fig. 3(b) is $\{d_3, d_5, d_7\}$.

2.2. High Recall Retrieval on a graph is NP-hard

Let us recall a couple of graph-related definitions. A *dominating set* $\mathcal{S}_D$ for a graph G is a subset of vertices of G such that every vertex of G not in $\mathcal{S}_D$ is joined with at least one vertex of $\mathcal{S}_D$ by some edges. An *independent set* $\mathcal{S}_I$ for a graph G is set of vertices of G such that for every two vertices in $\mathcal{S}_I$, there is no edge connecting them. It is clear that the dominating and the independent sets ($\mathcal{S}$) of G satisfy the coverage and the dissimilarity conditions, respectively, in definition ε-*ReS*. Hence the following observation is defined.

Observation 1. Solving the Minimum ε-*ReS* Diverse set problem for a set $\mathcal{D}$ is equivalent to finding the *Minimum Independent Dominating Set* of the corresponding graph G.

Definition 4 (*Minimum Independent Dominating Set***).** The *Minimum Independent Dominating Set* $\mathcal{S}^*$ is a subset of V from Graph G such that every vertex not in $\mathcal{D}$ is adjacent to at least one member of $\mathcal{D}$ and it holds that $|\mathcal{D}^*| \leq |\mathcal{D}|$.

The *Minimum Independent Dominating Set* is an NP-hard problem [4]. The problem remains NP-hard even for special kinds of graphs, such as *Unit Disk Graph* (UDG) [5]. *Unit Disk Graph* is a graph whose vertices can be put in one to one correspondence with equalized circles in a plane such that two vertices are joined by an edge, if and only if, the corresponding distances are intersected. Hence $G_{\mathcal{D},\varepsilon}$ is considered as a *Unit Disk Graph*, with respect to the Euclidean distance and this demonstrates that our problem is also NP-hard.

3. High Recall Retrieval with a Single-Step (HRR^1)

Before presenting our main HRR^2 (High Recall Retrieval with a Double-Step) method, we first describe a conceptual simple scheme for a better understanding. We call this scheme HRR^1 (High Recall Retrieval with a single-step) which works well for data in small size. We will then present HRR^2 in Section 4 which is applicable on large scale datasets.

The HRR^1 method (in Algorithm 1) consists of two stages. The first is to perform *Minimum ε-ReS Diverse set* within the entire collection which is the process of selecting a minimum number of documents to find k relevant documents (in Algorithm 2). The second is to determine the order of efficient examination and find k relevant documents (the line 6 in Algorithm 1). This process finds the promising region where relevant documents are gathered. We discuss the second stage in the next section and the first stage is described as follows.

Algorithm 2 is considered to compute the *Minimum ε-ReS Diverse set* $\mathcal{S}$ of G. For presentation convenience, let us call *green* the vertices of G that are in $\mathcal{S}$, *gray* the vertices covered by $\mathcal{S}$ and *white* the vertices that are neither *green* nor *gray*. Initially, $\mathcal{S}$ is empty and all vertices are white. The algorithm proceeds as follows. Up to the point that there are no more white vertices, it selects the largest degree of vertex, $\max(d(v_i))$ (ties may be broken arbitrarily), colors v_i *green* and colors all vertices in $N_\varepsilon(v_i)$ *gray*. We call this

Table 1
The example for algorithm description.

v_i	$v_j \in \mathcal{N}_\varepsilon(v_i)$	$\lvert\mathcal{N}_\varepsilon(v_i)\rvert$
v_3	$\{v_1, v_2, v_4, v_6\}$	4
v_5	$\{v_2, v_6, v_8, v_9\}$	4
v_2	$\{v_1, v_3, v_5\}$	3
v_6	$\{v_3, v_5, v_7\}$	3
v_1	$\{v_2, v_3\}$	2
v_4	$\{v_3\}$	1
v_7	$\{v_6\}$	1
v_8	$\{v_5\}$	1
v_9	$\{v_5\}$	1

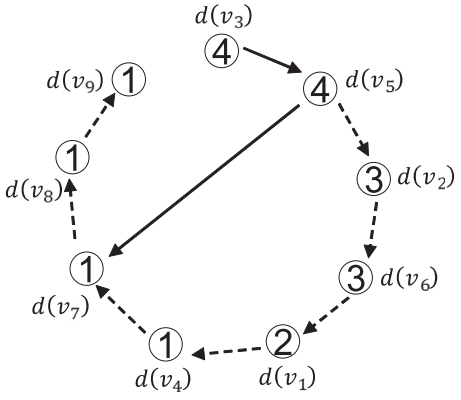


Fig. 4. The example of examination sequence by $|d(v_i)|$.

algorithm Representative Sampling, where $\mathcal{N}_\varepsilon^W(v_i)$ is the set of white neighbors of vertex v_i. Table 1 provides information on the example of Fig. 3(b) in which all vertices with their neighborhoods are presented in a descending order. For example, at first, v_3 is selected to be in $\mathcal{S}$ since it has the largest degree in G. If v_3 was selected to be in $\mathcal{S}$, then all the nodes (v_1, v_2, v_4, v_6) in v_3's neighborhood will become *gray*. In the next step, v_5 is selected as the largest node among the remained *white* nodes to be in $\mathcal{S}$. Eventually, v_7 is the last candidate to be in $\mathcal{S}$ and since there is no *white* node in G anymore, the algorithm is terminated with a *Minimum ε-ReS Diverse set* $\mathcal{S} = \{v_3, v_5, v_7\}$. As shown in Fig. 4, based on our assumptions, we review only three representative documents (v_3, v_5, v_7) and find all relevant documents (v_4, v_7).

Algorithm 1. HRR[1]: High Recall Retrieval with a Single-Step.

Input: a graph $G_{\mathcal{D},\varepsilon}$, integer k.
Output: k relevant documents set $\widetilde{\mathcal{D}}$.
1: $\widetilde{\mathcal{D}} \leftarrow \varnothing$
2: $\mathcal{U}_{Reviewed} \leftarrow \varnothing$
3: invoke Representative Sampling $\mathcal{S}$ to retrieve
4: **while** $|\widetilde{\mathcal{D}}|$ is k **do**
5: **for** $v_i \in \mathcal{S} \backslash \mathcal{U}_{Reviewed}$ **do**
6: $v_i^* \leftarrow \text{argmax}_{v \in \mathcal{S} \backslash \mathcal{U}} \text{Div}$
7: $\mathcal{U}_{Reviewed} \leftarrow v_i^*$
8: Review the documents $\mathcal{N}(v_i^*)$
9: **if** $\mathcal{N}(v_i^*)$ have the relevant documents **then**
10: $\widetilde{\mathcal{D}} \leftarrow \widetilde{\mathcal{D}} \oplus$relevant documents $\in \mathcal{N}(v_i^*)$
11: return $\widetilde{\mathcal{D}}$

Algorithm 2. Representative Sampling;*ReS*.

Input: A set of object $\mathcal{D}$ and radius ε.
Output: Representative Sampling $\mathcal{S}$ of $\mathcal{D}$
1: $\mathcal{S} \leftarrow \varnothing$

2: **for each** $v_i \in \mathcal{D}$ **do**
3: color v_i white
4: **while** there exist white objects **do**
5: select the white object v_i with the largest $|\mathcal{N}_\varepsilon^W(v_i)|$
6: $\mathcal{S} = \mathcal{S} \cup v_i$
7: color v_i green
8: **for each** $v_j \in \mathcal{N}_\varepsilon^W(v_i)$ **do**
9: color d_i gray
10: return $\mathcal{S}$

However, HRR[1] has the following limitations. In general, reducing the size of ε will increase the accuracy of the neighbor's information, but the average degree of the graph will be smaller. In the end, the number of *ReS* will increase and the review efforts by a user will increase (As shown in Fig. 5). This is a contradiction. To effectively solve this problem, we propose H*RR*[2] in the Section 4 after discuss the upper bound of our representative sampling strategy (*ReS*) for HRR[1].

3.1. The upper bound of Representative Sampling

The size of $\mathcal{S}$ (the cardinality of Representative Sampling) is affected by the value of ε. If the value of ε increases, the size of $\mathcal{S}$ decreases and vice verse. In this study, however, it is assumed that the document information within the $\mathcal{N}_\varepsilon(d_i)$. Another important issue to consider is finding an upper bound of representative sampling. This answers the question that if we examine a huge number of documents within a given ε, can we grasp the entire collection.

3.1.1. Dominating and independent dominating sets

A *dominating set* of a graph G is a set S_D of vertices of G such that every vertex not in S_D is adjacent to a vertex in S_D. The *domination number* of G, denoted by $|S_D|$, is the minimum size of a dominating set. A set is *independent* (or *stable*) if no two vertices in it are adjacent. An *independent dominating set* of G is a set that is both dominating and independent in G. The *independent domination number* of G, denoted by $|\mathcal{S}|$, is the minimum size of an independent dominating set. The *independence number* of G, denoted $|S_I|$, is the maximum size of an independent set in G. From the definitions, it follows immediately that $|S_D| \le |\mathcal{S}| \le |S_I|$. For example, in Fig. 3, S_D is $\{v_3, v_5, v_6\}$, $|S_D| = 3$, $\mathcal{S}$ is $\{v_3, v_5, v_7\}$, $|\mathcal{S}| = 3$, and S_I is $\{v_1, v_4, v_5, v_7\}$, $|S_I| = 4$.

3.1.2. Bounds on the independent domination number

The total number of documents to be checked is a very important issue which demonstrates the importance of the upper bound in our problem. In the first published study in this area, Berge [6] established a simple relationship between the independent domination number and the maximum degree of a graph. Later the upper bound was improved by Blidia et al [7]. Earlier, Bollobs and Cockayne [8] observed the following useful property of minimum dominating sets.

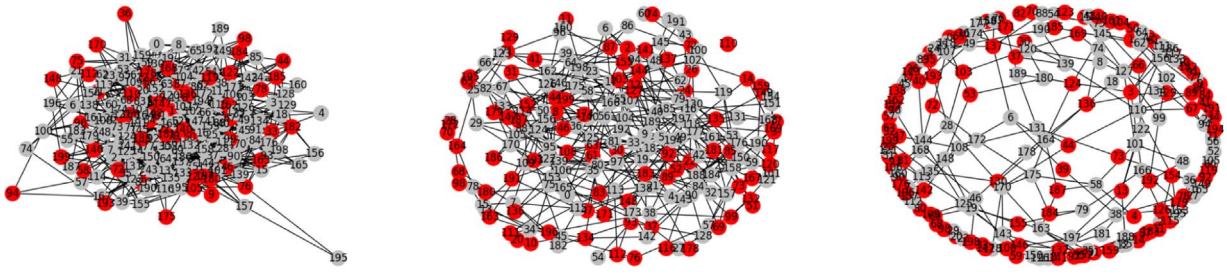
Proposition 1. *For a graph G with n vertices and maximum degree* Δ,

$$\left\lceil \frac{n}{1+\Delta} \right\rceil \le |\mathcal{S}| \le n - \Delta.$$

For example, Δ is 4 in Fig. 3, therefore $\lceil 9/(1+4) \rceil \le |\mathcal{S}| \le 9 - 4 \approx 2 \le |\mathcal{S}| \le 5$ using proposition 1.

Observation 2. If G is a connected graph, then there exist S_D such that for every $v \in S_D$, there exist a vertex $u \in V(G) \backslash S_D$ such that $\mathcal{N}[u] \cap S_D = v$ (called an external private neighbor).

Using this observation, Bollobás and Cockayne [8] proved the following upper bound on the independent domination number.



(a) Total of 55 ReS set of an random graph (800 Arcs)

(b) Total of 82 ReS set of an random graph (400 Arcs)

(c) Total of 100 ReS set of an random graph (200 Arcs)

Fig. 5. Representative Set of an random graph (The number of nodes is 200, the red color is *ReS*).

Theorem 1. *If G is a connected graph on n, then* $|\mathcal{S}| \leq n + 2 - |S_D| - \left\lceil \frac{n}{|S_D|} \right\rceil$

Proof. Regarding Observation 2, there exists a S_D such that every vertex $v \in S_D$ has an external private neighbor. For each vertex $v \in S_D$, choose an external private neighbor v'. By the Pigeonhole Principle, there is a vertex $y \in S_D$ that is adjacent to at least $(n - |S_D|)/|S_D|$ vertices of $V(G)\backslash S_D$. Let S'_D be a maximal independent set containing y. Since $S'_D \cap N(y) = 0$ and S'_D can contain at most one of x and x' for every vertex $x \in S_D\backslash y$, it follows that $|S'_D| \leq n - (|S_D| - 1) - \left\lceil \frac{(n - 1 - |S_D|)}{|S_D|} \right\rceil$. Since $\mathcal{S} \leq |S'_D|$, the result follows. □

Since the upper bound in Theorem 1 is maximized at $S_D = \sqrt{n}$, one immediately obtains the following bound, first noted by Favaron [9]:

Theorem 2. If G is a connected graph on n, then $|\mathcal{S}| \leq n + 2 - 2\sqrt{n}$

Now consider a graph on n vertices with a minimum degree of at least δ. Favaron [9] proved an upper bound on $i(G)$ for $\delta \geq 2$, and he conjectured the extremal value as a function of n and δ.

Theorem 3. *If G is a connected graph and has minimum degree at least δ, then* $|\mathcal{S}| \leq n + 2\delta - 2\sqrt{\delta n}$

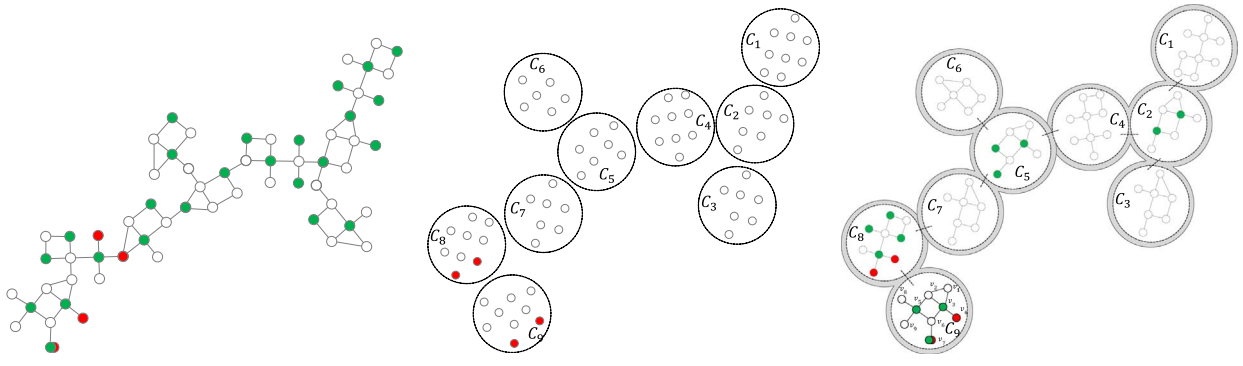
If the following conditions exist (n=5,000, $\delta = 2$), In the worst case, we need to check almost every document regardless of the number of k. To improve this, we should both consider effective documents partitioning strategy and efficient retrieval order. We will introduce the developed retrieval technique in the next section.

4. High Recall Retrieval with a Double-Step (HRR^2)

As mentioned in Section 3, HRR^1 is not applicable on large scale datasets. The size of *ReS* is often too large which brings about a very expensive review and collection process. To tackle this problem, we introduce the High Recall Retrieval with a Double-Step (HRR^2) that partitions a data graph into multiple clusters. Many clustering methods have been proposed in the literature (e.g., [10–13]). Since our main focus is on retrieval process and representative sampling, we will not delve into specifics here. Instead, we provide a general definition for the clustering function and discuss the features and properties relevant to our efficient retrieval process.

Definition 5. Given a result set of documents (vertices) $V = \{v_1, v_2, \ldots, v_n\}$, the process of partitioning V into $C = \{C_1, C_2, \ldots, C_K\}$ based on a certain distance measure, and C_i's are clusters, where $C_i \subseteq V$, $(i = 1, 2, \ldots, K)$, $\cap_{i=1}^{K} C_i = \varnothing$ and $\cup_{i=1}^{K} C_i = V$

However, clustering the entire dataset could not solve our problem. If we perform a representative sampling of all clusters to find the entire k related documents, we end up with the same utility (examination cost) result as HRR^1. To solve this problem, we present an efficient method for examination cluster selection based on diversity. For better understanding, the clarification of the process is provided with an example in Fig. 6. Assume that the whole 81 documents in the figure are clustered into nine clusters (C_1, …, C_9). First, select the cluster you want to examine first with a certain criterion. A good example is a cluster with a high cardinality or a cluster that contain the existing first ranking document. This example assumes that the cluster is C_2. Then perform *ReS* on cluster C_2. Therefore, the *Minimum ε-ReS Diverse set* is determined for C_2 and it is checked whether it includes any relevant documents or not. It should be noted that it is assumed we are aware of the state (relevant or non-relevant) of the documents (neighbors) which linked to any representative document. Hence, according to our assumption, it is found out that C_2 does not include any relevant document. Since C_2 is free of the relevant document, the process proceeds to go to the farthest cluster (C_8) among the rest of representative clusters (C_5 and C_8). Then, the *Minimum ε-ReS Diverse set* for C_8 is defined and after checking the neighbors of representative documents, two relevant documents are recognized. As C_8 contains two relevant documents, the closest cluster (C_9) to C_8 is selected to be checked for relevant documents. The *Minimum ε-ReS Diverse set* $\{v_3, v_5, v_7\}$ for C_9 is then determined and it is realized that the node v_7 which is a representative node itself and node v_4 which is the neighbor of another representative node (v_3) are relevant documents. According to the assumption that the number of relevant documents is given, the process is terminated since four relevant documents were detected. The process shows that our method found the relevant documents by checking only 9 documents



(a) näive High Recall Retrieval on G (b) The process of HRR^2 (c) The process of HRR^2

Fig. 6. The process of Effective High Recall Retrieval.

out of 81 documents as quickly and efficiently as possible. If C_8 is *ReS*, and there is no related document, then the cluster will select the cluster that is farthest from C_2 and C_8 as the next cluster. In the next section, we will present the equations and the algorithms of our method.

4.1. HRR^2 algorithm

An example of the examination sequence using Table 1, has been depicted in Fig. 4. As can be seen in the figure, the examination sequence is based on the descending order of nodes degrees. Hence, the process starts from checking node v_3 with a degree of 4. No matter v_3 is a relevant or non-relevant document, $v_5(d(v_5) = 4)$ is selected as the next document to be checked and the process keeps continuing up to the last documents v_9 and finally the process is finished. However, this method is inefficient if the results of the samples are gathered in a similar area. As a result, the much more time-consuming process will be needed. Thus, it is important to perform the examination sequence properly. The objective function of the method has been demonstrated in Eq. (2). It represents the maximization of re-ordering result O_S of $\mathcal{S}$ by, in the meantime, considering the nodes degrees and the diversity. Given $v_i \in \mathcal{S}$ and a variable w_i, $\frac{d(v_i)}{\sum_{i=1}^{|\mathcal{S}|} d(v_i)}$ represents the differentiation importance of sequence position for each v_i. But, this equation is no easy solution for directly optimizing it. So, we present an efficient document ordering algorithm by sequentially optimizing the objective function in Eq. (3).

$$\max O_{\mathcal{S}} = \sum_{i=1}^{|\mathcal{S}|} w_i \cdot d(v_i) + b \cdot \sum_{i=1}^{|\mathcal{S}|} \sum_{j=i+1}^{|\mathcal{S}|} dist(v_i, v_j) \tag{2}$$

$$O_t - O_{t-1} = d(v_t) + \sum_{i=1}^{t-1} dist(v_i, v_t) \tag{3}$$

The objective is to select a document that has a maximum increase of the objective function. Notice that such a sequential update may not necessarily provide an optimal solution, but it provides an excellent trade-off between accuracy and efficiency. The increase of the objective function from position $t-1$ to t is:

Algorithm 3 represents the pseudo code of the HRR^2 method for a dynamic high retrieval problem. The framework's goal is to minimize the burden of the time-consuming task by applying the macroscopic level (clusters level), and the microscopic level (documents level), where the macroscopic level is responsible for finding promising areas among the clusters quickly, and the microscopic level is in charge of finding the diverse representative sampling of documents within a cluster. Algorithm 3 first computes the representative clusters $\mathcal{S}_C$ by the given clustered result set C. Then, the algorithm recalculates the diversity re-ordering for reviewing sequence in $\mathcal{S}_C$ with $\max O_{\mathcal{S}}$ considering diversity, in line 4. Among all the clusters in $\mathcal{S}_C$, the algorithm picks the one that leads to the largest number of neighbors and maximum distance from target cluster. The line 7 of the algorithm is allocated to find the representative documents within the cluster in a similar fashion of clusters. The algorithm is terminated after finding the all k relevant documents.

Algorithm 3. HRR^2: High Recall Retrieval with a Double-Step

Input: The clustered result set C, integer k
Output: k relevant documents set $\widetilde{\mathcal{D}}$

1: $\widetilde{\mathcal{D}} \leftarrow \varnothing$;
2: pop the cluster C_i
3: $C_p \leftarrow C_i$
4: **while** there exist C or $|\widetilde{\mathcal{D}}|$ is k **do**
5: Representative Sampling $\mathcal{S}$ on C_p
6: Review the documents $\mathcal{S}$
7: **if** $\mathcal{S}$ have the relevant documents **then**
8: $\widetilde{\mathcal{D}} \leftarrow \widetilde{\mathcal{D}} \oplus$relevant documents on $\mathcal{S}$
9: $C_i^* = \arg\min_{C_i \in N(C_p)} dist(C_i, C_p)$
10: $C \leftarrow C \backslash C_i$
11: $C_p \leftarrow C_i^*$
12: **else**
13: $C_i^* = \arg\max_{C_i \in \mathcal{S}_C} dist(C_i, C_p)$
14: $C \leftarrow C \backslash C_i$
15: $C_p \leftarrow C_i^*$
16: **return** $\widetilde{\mathcal{D}}$

Table 2
The performance of the different scores with different IR systems.

SYSTEM	Ranks of rel. docs	*AP*	F_1	F'_4	*PRES*	review efforts
1	{1, 2, 3, 4}	1	0.08	1	1	4
2	{49, 50, 51, 52}	0.05	0.08	0.47	0.52	52
3	{1, 98, 99, 100}	0.27	0.08	0.86	0.28	100
4	{97, 98, 99, 100}	0.03	0.08	0.31	0.04	100

4.2. Methodology analysis

Theorem 4. *If* $\lceil\sqrt{n}\rceil$ *number of clusters are considered, then the maximum number of documents to be searched with our Dynamic Diversity Retrieval strategy is* $|\mathcal{S}| \leq (\lceil\sqrt{n}\rceil + 2\delta - 2\lceil\sqrt[4]{n\delta^2}\rceil)^2$.

For example, in Fig. 6, n=81 and there are 9 clusters ($|C| = \lceil\sqrt{81}\rceil$), so the maximum number of documents to be checked is 25.

5. Evaluation metrics

The simplest evaluation measure to assess the retrieval performance is evaluating the recall. However, the problem of doing this is that it fails to reflect how early a system retrieves the relevant documents, and thus a number of user review efforts that can not be counted. Table 2 shows an illustrative example of how different metrics perform with four different IR systems when a collection is searched by a given query. In this case, there are four relevant document results, and it is assumed that the user is willing to check the top 100 documents retrieved for finding all relevant documents by each of the four systems. We can compare the following four representative evaluation metrics. The first measure is Average Precision(*AP*) [1], the most popularly used metric. The second one is $F_1 - Measure$, which has been one of the solutions to measure the performance assessment in the recall focused information retrieval task. The third one is $F'_4 - Measure$ [14], which is similar to the $F_1 - Measure$ but heavily weighted on *recall*. The last one is *PRES* (Patent Retrieval Evaluation Score) [15], a recent recall-oriented evaluation metric, where the higher the *PRES* value, the lower the user cost to find all the relevant documents.

As an example, the user in system 1 finds all relevant documents (k=4) after checking the ranked list up to the fourth document. This is the best case so that the *AP*, F'_4, and the *PRES* value are equal to 1. However, in a case of system 2, the relevant documents are ranked in the middle of the results between 49 and 52, so the user will find all relevant documents after checking the ranked list up to 52, and the *PRES* value here is equal to 0.52. This demonstrates that the *PRES* value is one of the recall-based evaluation methods that consider the user retrieval cost. Therefore, we use the *RRES*, which best reflects the review efforts, as an evaluation metric.

6. Experiment results

In this section, we have presented empirical experiments to evaluate the effectiveness of the HRR^2 compared with ReQ-ReC (ReQuery-ReClassify) and RF (Relevance Feedback). The ReQ-ReC [3] is the state-of-the-art query expansion (Rocchio) with SVM. The RF is Rocchio relevance feedback method [1]. Two data sets (Yeast and 20-newsgroup) were used in the evaluation. The criterion for the selection fo suitable datasets is that the whole data set must have relevant information and this is because our goal is to achieve a 100% recall. Table 3 shows the total number of documents ($|\mathcal{D}|$), attributes ($|m|$), and classes ($|class|$) related to each dataset. We have also used a high recall retrieval metric called *PRES* [15], which depends more on the recall. Note that many popular metrics such as *precision@k*, *MAP*, and *nDCG* exist for retrieval performance but they not appropriate for high recall tasks nevertheless.

6.1. Yeast data set experiment

This section provides the experimentation result of Yeast dataset. Table 4(a) illustrates the different classes and the number of relevant documents related to each class in Yeast dataset. As can be seen in the table, Yeast dataset consists of 10 different classes with at least 5 and at most 463 relevant items in each class.

According to Table 4, for example, the class ‘ERL’ has 5 relevant documents among total 1484 items. As shown in Fig. 7, the RF method in this case, has checked 15 items to find the all 5 relevant ones, whereas ReQ-ReC and HRR^2 methods have quested 141 and 6 items, respectively. In another case for instance ‘ME1’ including 44 relevant items, the RF, ReQ-ReC, and HRR^2 methods have read

Table 3
The Datasets.

Name	$\|\mathcal{D}\|$	$\|m\|$	$\|Class\|$
Yeast	1,484	8	10
20NG	18,774	61,188	20

Table 4
The experiment of Yeast Dataset.

CLASS (n = 1, 484)	k
ERL (endoplasmic reticulum lumen)	5
POX (peroxisomal)	20
VAC (vacuolar)	30
EXC (extracellular)	35
ME1 (membrane protein, cleaved signal)	44
ME2 (membrane protein, uncleaved signal)	51
ME3 (membrane protein, no N-terminal signal)	163
MIT (mitochondrial)	244
NUC (nuclear)	429
CYT (cytosolic or cytoskeletal)	463

(a) Yeast Dataset Introduction

	RF	ReQ-ReC	HRR^2
ERL	0.997	0.982	**1.000**
POX	0.775	0.707	**0.998**
VAC	0.469	0.470	**0.599**
EXC	**0.845**	0.785	0.821
ME1	0.889	0.928	**0.942**
ME2	0.760	0.792	**0.894**
ME3	0.302	0.718	**0.831**
MIT	**0.754**	0.693	**0.754**
NUC	0.514	0.602	**0.807**
CYT	0.512	0.501	**0.740**

(b) The Comparison of PRES on Yeast Dataset

629, 517 and 197 items to find the all 44 relevant items, respectively. Likewise, the equalization method outperforms the other approaches in terms of *PRES* values, as illustrated in Table 4(b). It can be clearly seen in the table that all the *PRES* balues except 'EXC' case in HRR^2column are higher than the RF and ReQ-ReC ones. The experimentation results for each class have been also depicted in Fig. 7 separately. As shown in the figure, even in the worst case 'CYT', the HRR^2 method has checked 1041 items to find the all 463 relevant ones and this shows that the performance of our method even in the worst case has been 1.5 times better than the others. As a conclusion, in all cases, our method outperforms the ReQ-ReC method as well as the conventional one (RF).

6.2. 20NG set experiment

In the second experiment, the 20 Newsgroups(20NG) dataset is used, which data set is a collection of 18,774 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. As shown in Table 5, 20NG dataset consists of 20 different classes of at least 627 and at most 997 relevant items in each class. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). We have cinsidered two different classes ('comp.sys.mac.hardware' and 'misc.forsale') in our experiment to evaluate the comparison of the *PRES* values among RF, ReQ-ReC, and HRR^2 methods. Generally, it is not easily achievable to find the whole relevant documents from 'comp.sys.mac.hardware', because this class is closely related to some other classes. On the other hand, finding the whole relevant document in 'misc.forsale' is relatively easy since it does not have any relation with other classes. Table 6 provides information on the comparison of *PRES* values. The results show that our algorithm has higher values of *PRES* in both classes and therefore outperforms the others.

7. Related work

The research of patent information retrieval is mainly divided into patent search and patent analysis. The first, *patent search* is concerned with finding all filed patents relevant to a given patent application. The queries in patent retrieval are typically very long since they take the form of a patent claim or even a full patent application in the case of prior-art patent search. These types of research are called query formulation and query expansion [16–18].

The second is *patent analysis* which finds the semantic information by analyzing the relevant patent results set. The patent analysis techniques have mainly been exploited by text mining and visualization techniques. The text mining techniques further utilize Natural Language Processing (NLP) approaches, semantic analysis approaches, rules-based approaches, property function approaches, and neural networks approaches [19–22]. On the other hand, visualization techniques for patent analysis also use certain text mining methods to present the results of patents in visual forms. The visual output task of the patent analysis is in the form of patent networks, patent maps, and data clustering [23,24,13].

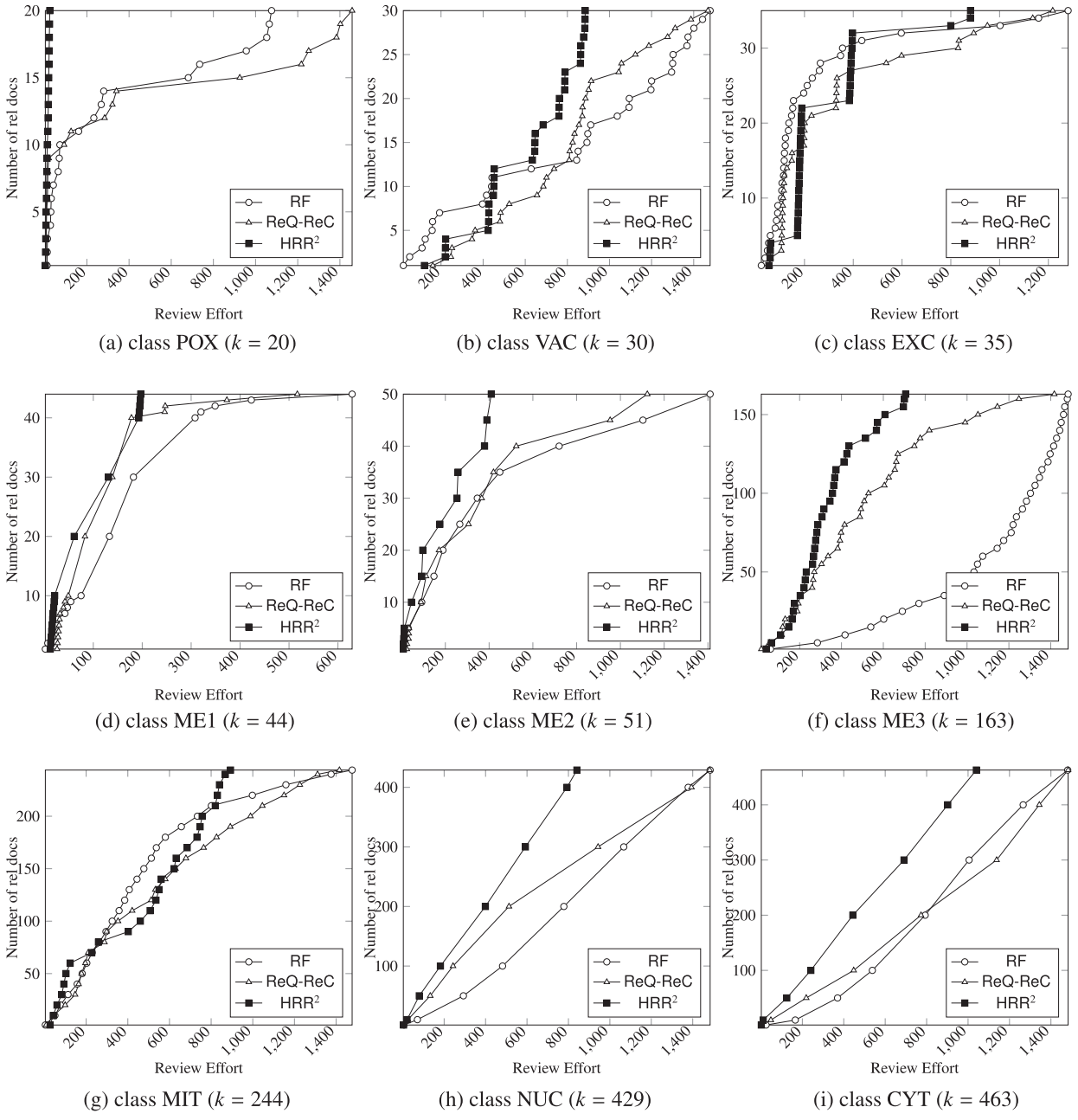


Fig. 7. Yeast Data Sets Experiments (POX, VAC, EXC, ME1, ME2, ME3, MIT, NUC, and CYT).

7.1. Query expansion techniques

Many query expansion techniques have been introduced in the field of IR with the goal of improving retrieval effectiveness. The primary objective of query expansion is to overcome the mismatch between search queries and relevant documents. This is a typical situation that queries are short and do not describe enough the users information needs well. However, the query is typically very long for the patent retrieval while there is still often a significant mismatch between queries and relevant items.

Relevance feedback has long been suggested as an effective method for improving retrieval performance [25]. In a feedback procedure, the retrieval system presents the top-ranked documents to the user and collects back either explicit judgments of these documents or implicit feedback implied by certain actions of the user [26]. The system then learns from the collected feedback and updates the query. The new query indicates a refined understanding of the user's information needs, which enhances both precision and recall in the next retrieval step. Even without actual user judgments, retrieval performance may still benefit from directly employing the top-ranked documents as a relevant input, which is known as a process of pseudo-relevance feedback [25,27]. When it

Table 5
20NG Dataset Introduction.

label	label name	*k*	label	label name	*k*
1	alt.atheism	798	11	rec.sport.hockey	997
2	comp.graphics	970	12	sci.crypt	989
3	comp.os.ms-windows.misc	963	13	sci.electronics	984
4	comp.sys.ibm.pc.hardware	979	14	sci.med	987
5	comp.sys.mac.hardware	958	15	sci.space	985
6	comp.windows.x	982	16	soc.religion.christian	997
7	misc.forsale	964	17	talk.politics.guns	909
8	rec.autos	987	18	talk.politics.mideast	940
9	rec.motorcycles	993	19	talk.politics.misc	774
10	rec.sport.baseball	991	20	talk.religion.misc	627

is possible to process the entire collection of documents, the problem of high-recall retrieval can be phrased as a binary classification problem where the positive class captures documents that are relevant to the information needs and the negative class captures the non-relevant ones. The practice of relevance feedback radically accelerates the active learning process, in which the method interactively accumulates the training result by selecting the relevant documents and requesting the user for giving labels.

These aforementioned goals have led researchers to investigate query expansion techniques for patent search. However, reported work on query expansion for patent search has never demonstrated consistent effectiveness [16]. Some of the initial trials utilizing Pseudo Relevance Feedback (PRF) for query expansion in patent search are described in [28]. PRF is a standard technique used to improve a search query with additional terms from the top ranked documents based on an initial retrieval run under the assumption that these documents are relevant [25]. In this work, a novel mechanism for PRF specifically designed for the patent search was introduced and compared to the standard Rocchio method. Experiments on the NTCIR-3 (NII Testbeds and Community for Information access Research) patent retrieval task did not produce any significant improvement in retrieval results. The author commented the reason for this might be that all words from the documents assumed to be relevant were used without any selection process. In NTCIR-4, there was another attempt at utilizing Query Expansion (QE) through PRF to improve the retrieval effectiveness [29]. However, it was found that while retrieval effectiveness was improved for a few topics, it was degraded for many others. The authors did not provide a clear analysis of possible reasons.

7.2. Patent analysis techniques

7.2.1. Text mining techniques

Text mining is a knowledge-based process that uses analytic tools to derive meaningful information from Natural Language Processing (NLP). The information is derived from the text by identifying and detecting significant patterns from unknown textual data. NLP is a text mining approach that uses computational mechanisms to analyze and represent the textual information quantitatively as well as semantically. In patent analysis, NLP has also been utilized for the transformation of technological information into simple language structures by extracting the grammatical analysis from the textual data and creating the structural relationships among the components [20,21,30,31]. However, the NLP-based approaches on patent analysis have suffered from the issues of lexical and grammatical ambiguities and also lack in representing the semantic relationships among the grammatical structures. The property-function analysis approach extracts properties and functions from patent documents as innovating concepts through grammatical analysis. The property expresses a particular characteristic of a system whereas the property function represents a suitable action of the system [22]. Unlike keyword approaches, property-function based methods do not require the re-defined set of keywords and key phrase patterns. Despite their usefulness, the property-function based techniques have also exposed similar drawbacks to other text mining and NLP based techniques. Rule-based techniques for text mining have mostly used some inference and association rules. Such kinds of techniques are effective for creating meaningful associations among the structures extracted from large data sets [30]. The rules are usually an IF-THEN procedure that helps in extracting the appropriate data from the patents. However, the rule-based approaches have limitations, since the instance rules exhibit incompetence in representing the incomplete knowledge. Moreover, as the number of the rules in the rule bases increases, the risk of obtaining spurious associations among the rules also increases [32]. Semantic-based text mining techniques rely on the domain knowledge and create a relationship among the domain specific concepts [31]. The types of techniques are effective in identifying the similarities among patents and determining the future technological trends by logically relating parsed grammatical structures. However, semantic-based approaches have faced problems particular to parsing the structures of natural language. Therefore, the semantic analysis based

Table 6
The result of PRES on the 20NG.

	RF	ReQ-ReC	HRR^2
comp.sys.mac.hardware	0.432	0.456	**0.560**
misc.forsale	0.235	0.765	**0.891**

approaches could exhibit the inadequacy in accurately representing the concepts. Neural network based approaches have also been used for patent classification and technology forecasting [33]. More specifically, the back propagation in the neural network algorithm has been used to train a patent network to determine the quality of patents. However, the approach may suffer from the cold-start problem.

7.2.2. Visualization techniques

Another major approach for contemporary patent analysis is the use of visualization tools to represent patent information and the result analyses. Chang et al. [23] presented a framework to the technology trends identification using the patent network, bibliometric patent analysis with graphs and quantitative technique for constructing networks. Tang et al. [34] proposed a model called Inventor-Company-Topic (ICT) model that incorporates information about the inventors and companies using the patent network generation. Kim et al. [35] presented a visualization method for patent analysis that is to determine the trend shift for certain technologies using the *k*-means clustering method from the semantic network of keywords to determine meaningful relationships. Segev and Kantola [36] developed a model to identify new research directions with self-organizing maps from the extraction of patents terms, context retrieval, and context ranking. Although visualization techniques represent the information extracted from the patents, they still depend on the text mining approaches to extract the information from patent documents. The visualization techniques using text mining approaches have suffered from the similar limitations of the text mining approaches.

8. Conclusion

We presented algorithms named HRR^1 and HRR^2 which are suitable for the high recall retrieval problem without sacrificing precision and minimize the review efforts. We also presented the theoretically proved that our approach can reduce the upper bound effectively. Given a certain *precision*, the efforts to achieve the full *recall* level can be processed so that the HRR^1 and HRR^2 algorithms will find the most promising region and decide to move to the next promising region dynamically based on the double-step independent domination set strategy. By applying the proposed method, the user's efforts can significantly be reduced. The various datasets were used in our experimentation and all the results demonstrated that our method outperforms the ReQ-ReC as well as the conventional method (RF).

For the future work, our method can also be useful for other applications such as making a patent map, patent portfolio, investment strategy, etc. Also, the optimal value of ε with an unknown number of relevant documents(k) can be specified later.

Acknowledgment

This work was supported by Inha University.

References

[1] R.A. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
[2] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv 34 (1) (2002) 1–47.
[3] C. Li, Y. Wang, P. Resnick, Q. Mei, Req-rec: High Recall Retrieval with Query Pooling and Interactive Classification, SIGIR '14, ACM, New York, NY, USA, 2014, pp. 163–172.
[4] M.R. Garey, D.S. Johnson, Computers and Intractability; A Guide to the Theory of NP-Completeness, W.H. Freeman & Co, New York, NY, USA, 1990.
[5] B.N. Clark, C.J. Colbourn, D.S. Johnson, Unit disk graphs, Discret. Math. 86 (1–3) (1990) 165–177.
[6] C. Berge, E. Minieka, Graphs and hypergraphs, Vol. 7, North-Holland publishing company Amsterdam, 1973.
[7] M. Blidia, A. Bouchou, L. Volkmann, Bounds on the k-independence and k-chromatic numbers of graphs, Ars Comb. 113 (2014) 33–46.
[8] B. Bollobás, E.J. Cockayne, Graph-theoretic parameters concerning domination, independence, and irredundance, J. Graph Theory 3 (3) (1979) 241–249.
[9] O. Favaron, Two relations between the parameters of independence and irredundance, Discret. Math. 70 (1) (1988) 17–20.
[10] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay, Clustering large graphs via the singular value decomposition, Mach. Learn. 56 (1–3) (2004) 9–33.
[11] W.B. Croft, Cluster-based retrieval using language models, Inf. Retriev. (2004) 186–193.
[12] N. Jardine, C.J. van Rijsbergen, The use of hierarchic clustering in information retrieval, Inf. Storage Retriev. 7 (5) (1971) 217–240.
[13] C. Shi, Y. Cai, D. Fu, Y. Dong, B. Wu, A link clustering based overlapping community detection algorithm, Data Knowl. Eng. 87 (2013) 394–404.
[14] C.J.V. Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, Newton, MA, USA, 1979.
[15] W. Magdy, G.J. Jones, Pres: A Score Metric for Evaluating Recall-oriented Information Retrieval Applications, SIGIR '10, ACM, New York, NY, USA, 2010, pp. 611–618.
[16] W. Magdy, G.J. Jones, A study on query expansion methods for patent retrieval, in: Proceedings of the 4th workshop on Patent information retrieval - PaIR '11, 2011, 19.
[17] W. Magdy, P. Lopez, G.J.F. Jones, Simple vs. sophisticated approaches for patent prior-art search, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 725–728.
[18] W. Magdy, J. Leveling, G.J.F. Jones, Exploring structured documents and query formulation techniques for patent retrieval, Lecture Notes in Computer Science 6241 LNCS, 2010, pp. 410–417.
[19] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, Inf. Process. Manag. 43 (5) (2007) 1216–1247.
[20] P. Masiakowski, S. Wang, Integration of software tools in patent analysis, World Patent Inf. 35 (2) (2013) 97–104.
[21] J. Yoon, H. Park, K. Kim, Identifying technological competition trends for r & d planning using dynamic patent maps: Sao-based content analysis, Scientometrics 94 (1) (2013) 313–331.
[22] T. Fleiner, Z. Jankó, Choice function-based two-sided markets: stability, lattice property, path independence and algorithms, Algorithms 7 (1) (2014) 32–59.
[23] P.-L. Chang, C.-C. Wu, H.-J. Leu, Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display, Scientometrics 82 (1) (2010) 5–19.
[24] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, A.K. Usadi, Patentminer: Topic-driven patent analysis and mining, KDD, ACM, 2012, pp. 1366–1374.
[25] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, J. Am. Soc. Inf. Sci. 41 (1990) 288–297.

[26] T. Joachims, L. Granka, B. Pan, H. Hembrooke, G. Gay, Accurately Interpreting Clickthrough Data as Implicit Feedback, SIGIR '05, ACM, New York, NY, USA, 2005, pp. 154–161.
[27] S. Yu, D. Cai, J.-R. Wen, W.-Y. Ma, Improving pseudo-relevance feedback in web information retrieval using web page segmentation, in: WWW, 2003, pp. 11–18.
[28] K. Kishida, Pseudo relevance feedback method based on taylor expansion of retrieval function in ntcir-3 patent retrieval task, PATENT '03, 2003, pp. 33–40.
[29] M. Lupu, K. Mayer, J. Tait, A.J. Trippe, Current challenges in patent information retrieval, Vol. 29, Springer-Verlag Berlin Heidelberg, 2011.
[30] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, SIGMOD 22 (2) (1993) 207–216.
[31] D. Bonino, A. Ciaramella, F. Corno, Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics, World Patent Inf. 32 (1) (2010) 30–38.
[32] G.D.F. Morales, A. Gionis, Streaming similarity self-join, PVLDB 9 (10) (2016) 792–803.
[33] H.T. Kahraman, A novel and powerful hybrid classifier method: development and testing of heuristic k-nn algorithm with fuzzy distance metric, Data Knowl. Eng. 103 (2012) 44–59.
[34] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, et al., Patentminer: topic-driven patent analysis and mining, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1366–1374.
[35] Y.G. Kim, J.H. Suh, S.C. Park, Visualization of patent analysis for emerging technology, Expert Syst. Appl. 34 (3) (2008) 1804–1812.
[36] A. Segev, J. Kantola, Identification of trends from patents using self-organizing maps, Expert Syst. Appl. 39 (18) (2012) 13235–13242.

Justin JongSu Song received his B.Sc. and M.Sc. in Industrial Engineering, with high honors, from Inha University, Korea, in 2012. He is currently Ph.D. candidate in Inha University. His research interests include Team Formation Problem, Social network, Information Retrieval, and Patent Analysis.

Wookey Lee received the B.S., M.S., and Ph.D. from Seoul National University, Korea, and the M.S.E. degree from Carnegie Mellon University, USA. He currently is a Professor in Inha University, Korea. He has served as chairs and PC members for many conferences such as CIKM, DASFAA, IEEE DEST, VLDB, BigComp, EDB, etc. He is currently one of the Executive Committee members of IEEE TCDE. He won the best paper awards in IEEE TCSC, KORMS and KIISE. Now he is the EIC of Journal of Information Technology and Architecture, and an associate editor for WWW Journal. His research interests include Cyber-Physical systems, Graph and Mobile systems, Data Anonymization, and Patent Information.

Jafar Afshar was born in Tehran, Iran in 1985. He received the B.Sc. (2009) and M.Sc. (2014) degrees in Industrial Engineering from Azad University, Iran, and Universiti Teknologi Malaysia, Malaysia, respectively. He is currently PhD candidate in Inha University. His research interests lie in Team Formation Problem, Information Retrieval, Patent Analysis, and Social Network.