Knowledge-Based Systems 163 (2019) 736–748

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# Dual pattern-enhanced representations model for query-focused multi-document summarisation

Yutong Wu, Yuefeng Li *, Yue Xu

*Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Australia*

ARTICLE INFO

*Article history:*
Received 17 January 2018
Received in revised form 20 September 2018
Accepted 25 September 2018
Available online 12 October 2018

*Keywords:*
Query-focused multi-document summarisation
Pattern mining
Topic modelling
Query expansion
Three-way decision theory
Unsupervised approach

ABSTRACT

To address the problem of query-focused multi-document summarisation, we present a novel unsupervised pattern-enhanced approach for representing coherent topics across documents, as well as the query relevance, in order to generate topically coherent summaries that meet the information needs of users. The proposed model employs not only a pattern-enhanced topic model to generate discriminative and semantic rich representations for topics and documents, but also a pattern-based relevance model for the query relevance of sentences. With these dual pattern-based representations for sentences, we are able to integrate various indicative metrics, such as rational coverage of document topics and sentence relevance, into a unified model. When evaluated on the datasets of the document understanding conferences of 2006 and 2007, the proposed approach shows a performance improvement as compared to a number of state-of-the-art methods and unsupervised baseline systems.



## 1. Introduction

Because it addresses the problem of information overload in the big data era, automatic multi-document summarisation (MDS) has increasing value for various real-world applications. MDS summarises information from multiple documents that share an explicit or implicit main topic. It helps users quickly catch the most relevant and important information in large text data collections. MDS can be categorised into generic [1,2] and query-focused MDS [3,4]. Generic MDS extracts the important content from a document collection without using any prior knowledge [5] or additional information, whereas query-focused MDS is aimed to produce a most typical summary reflecting the condensed information that is closely related to the initial given query, which expresses the information need of the user [6–8].

As a special case of MDS, query-focused summarisation has to handle both query relevance and topic coverage in a balanced manner [7], so that the user's information needs, including not only the answer to the given query but also the important information that the user requires, can be fulfilled within a text snippet that is limited in length. To address these challenges, the creation of an intermediate representation [9] for both documents and query so that the key aspects of both can be captured and measured is extremely important.

However, if we consider the topics of a given document set as the implicit query together with the given explicit query, then this creation of intermediate representation becomes a query expansion problem. From the viewpoint of query expansion in the context of query-focused MDS, most state-of-the-art approaches expand the key topics and query in a term-based manner, i.e., their objective is to represent the topics and query as a set of expanded relevant terms. These term-based methods are usually effective for query relevance, as it is simple and easy to measure the query relevance for sentences that contain some query-relevant terms. However, for sentences that contain no original query terms but may contain implicit information that is relevant or topically important, query expansion can become more complex and challenging, and sometimes external or hand-crafted resources, such as WordNet, labelled learning data, or Web data collections, have to be used to generate an accurate representation. Furthermore, a term-based representation, which treats a sentence or document as a bag of words, can attain only the literal or syntactic information contained in text documents, and ignores the semantic meanings, misses the contextual information of words or sentences and limits the performance of lexical cohesion discovery. Therefore, we need novel query expansion techniques for query-focused MDS tasks to expand explicit but short query terms and implicit topical terms, as well to attain a longer, more accurate, more coherent, more discriminative and more representative representation of the query and topics.

In this study, we employed pattern-based representation for both documents and query to overcome the limitations of term-based representation for query-focused MDS tasks. Since pattern mining, one of the most important and well-established techniques in the data mining area [10–13], can reveal interesting

* Corresponding author.
*E-mail address:* y2.li@qut.edu.au (Y. Li).

https://doi.org/10.1016/j.knosys.2018.09.035
0950-7051/

relations between data examples and discover multiple-word patterns, which usually carry more semantic contents and represent more specific meanings than single words, and requires no additional resources, we believe that its use, together with other query expansion techniques used in query-focused MDS, can help create more powerful representations for both explicit queries and implicit topics. To represent topics, we can utilise a pattern-based topic model [14,15] that combines pattern mining techniques with topic modelling to automatically generate discriminative and semantic rich representations for topics and documents and does not require any external resources. However, for query relevance representation, how to use a pattern-based model to represent and measure the query relevance of sentences for query-focused MDS remains a challenging problem. In the proposed query-focused summariser, instead of using a similarity-based relevance model, we create query relevant patterns to represent the query relevance of sentences. These query relevant patterns are derived from the sentence space, which consists of query-relevant sentences and background sentences, rather than the term space. In addition, a novel re-weighting algorithm based on three-way decision theory is applied to the discovered query relevant patterns through punishing some patterns that are not biased towards the given query in order to make our pattern-enhanced relevance model more discriminative and representative. On the basis of these dual pattern-enhanced representations, we are able to score sentences effectively and efficiently in terms of both their query relevance and their topic coverage and significance, and a moderate balance between query relevance and topics coverage can be easily maintained by a simple linear combination of these two scores. Thus, the most informative and relevant sentences can be selected to form a summary that answers the given query and also preserves the maximal information coverage of document contents.

The results of the study contribute to the field of query-focused MDS in the following perspectives: (i) the integration of two different but complementary pattern-based representations for query relevance and topic importance, (ii) the creation of a novel unsupervised pattern-enhanced query relevance model to represent the query relevance of sentences with the novel three-way decision enhanced *query relevant patterns*, and (iii) the creation of a set of accurate measures of sentences' topic significance, information coverage and query relevance for the proposed dual pattern-enhanced representations model.

The experimental results for the summarisation benchmark datasets, DUC 2006 and DUC 2007, both of which are open benchmarks datasets from the National Institute of Standards and Technology's (NIST) document understanding conferences[1] demonstrate that our proposed approach is effective and outperforms other unsupervised extraction-based query-focused MDS methods and many state-of-the-art systems.

In this study, we extended our previous work [15,16] on generic MDS to query-focused MDS tasks as follows. (i) We integrated a novel pattern-enhanced query-relevance representation with the pattern-enhanced topic representation for both topic importance and query relevance. (ii) We applied a novel re-weighting algorithm to query relevant patterns to reward discriminative patterns and punish patterns that are not biased towards the given query; the novelty of this algorithm can even contribute to other fields, such as information retrieval, in which the measurement of query relevance is also an important factor. (iii) We optimised sentence ranking algorithms with new or updated measuring methods. On the basis of these fundamental differences between this and our previous work, we are confident the approach proposed in this paper will provide new contributions to the fields of not only query-focused MDS but also information retrieval.

1 http://www-nlpir.nist.gov/projects/duc/index.html

## 2. Related work

As a challenging issue for text mining, automatic document summarisation had been studied in depth for 60 years [17]. The great majority of the numerous approaches developed are extraction-based [18,19], which produce a summary using only existing sentences (or text fragments) extracted from the original text, and thus, are conceptually simple and more practicable than abstractive methods, which attempt to reproduce sentences by using complicated natural language generation techniques, such as sentence compression [20], information fusion [21] and reformulation [22,23]. According to whether or not they require training samples, such as human-made summaries, summarisation methods also can be categorised into supervised or unsupervised methods. In this paper, we restrict the discussion to unsupervised, extractive, query-focused MDS.

Extraction-based query-focused summarisation methods need to select sentences to answer the query, as well as to yield salient topics coverage. Such models in general rely on different sentence ranking techniques, where the systems need first to establish an intermediate representation of the documents and/or query and then, based on this representation, to identify important content, score sentences and select top-ranked sentences to form the summary [9]. Some widely used summarisation methods rely on topic representations, where a variety of topic modelling techniques [24–29] are employed to generate topic representation for document summarisation. In recent years, the use of Bayesian topic models, using structured probabilistic topic models to generate an explicit representation for each individual document [30–33], has become increasingly widespread. While the statistics-based approaches consider only shallow features of documents, topic models also take into account semantic associations, so that they can capture information that is missed in most of the other approaches. However, since they focus only to a small extent on redundancy and coherence issues, topic representation approaches usually ignore the contextual information of words. To overcome this issue, Yang et al. [34] proposed a Bayesian hierarchical topic model that captures both the hierarchies and the word dependencies over latent topics by means of training a large data corpus before producing summaries of documents that have context features embedded systemically. Our model can use the advantages of topical models to complement a query-relevance model that may have a limitation in terms of implicit topic discovery for query-focused MDS tasks.

Pattern mining has been widely exploited to solve problems in the data mining [35,36], web intelligence and information retrieval (IR) communities [11,12,37] for many years. As compared to term-based methods, multi-word patterns that use the advantage of phrases are more discriminative and carry more semantics [38,39]. Some mature techniques, such as maximal patterns, closed patterns and master patterns, have been developed to provide a good alternative to phrases with reduced redundant and noisy patterns [40]. Gao et al. [14,41] combined pattern mining techniques with latent Dirichlet allocation (LDA) topic modelling to generate more discriminative and semantic rich topic representations by using patterns rather than single words. It also had been proved that the pattern-based key-phrase extraction method is effective in improving the quality of extracted key-phrases capturing a main topic of the underlying document [42]. However, in very few studies have successful attempts been made [43,44] to explore pattern mining techniques to solve the problem of automatic document summarisation. In ItemSum [43], which was the first attempt to exploit frequent patterns in MDS, the relevance score of a sentence is computed from the traditional bag-of-word (BOW) representation and the sentence coverage is validated in the measure of the pertinence of each sentence to a pattern-based model composed of

**Table 1**
Examples of word–topic assignments.

| Topic | Words |
|---|---|
| $Z_1$ | $w_4, w_5, w_7, w_8, w_{10}, w_{11}, w_{15}$ |
| $Z_2$ | $w_1, w_2, w_5, w_6, w_9, w_{12}, w_{13}, w_{15}$ |
| $Z_3$ | $w_1, w_2, w_3, w_5, w_6, w_7, w_{10}, w_{11}, w_{12}, w_{14}$ |

frequent patterns extracted from a transactional data format of the document collections. Unlike ItemSum, which exploits frequent patterns that are not sufficiently concise or accurate to represent specific topics that suffer the low frequency problem [37], PatSum, proposed by Qiang et al. [44], adopts frequent closed pattern for generic MDS. PatSum improves the performance of pattern-based models by removing the redundant patterns and solving the low frequency problem; however, the authors still ignored the issue of topical coherence of the sentences in MDS. In addition, how to adapt the closed patterns for an effective measure of query relevance, as well as of topic coverage, in query-focused MDS remains an open problem.

In this study, paying particular attention to addressing the aforementioned problems, we developed an effective pattern-based approach for query-focused summarisation by combining pattern-mining, topic modelling and query expansion techniques.

## 3. Dual pattern-enhanced representations for query-focused multi-document summarisation

### 3.1. Overview

The proposed model, namely dual pattern-enhanced representations for query-focused MDS (DPRQSum) employs dual pattern-based models to represent the topic importance and query relevance of sentences respectively in a given multi-document collection. As an extraction-based summariser, DPRQSum performs three typical tasks [9]: it creates an intermediate representation of document input, scores sentences based on the representation and selects sentences to form a summary. For query-focused summarisation, the extracted sentences must not only be relevant to the given query but also contain the most significant information in documents. Therefore, by applying pattern-enhanced query expansion techniques, DPRQSum generates two intermediate representations to lay a foundation for the two requirements of query-focused summarisation, one for query relevance and one for topics coverage. The architecture of our approach is shown in Fig. 1.

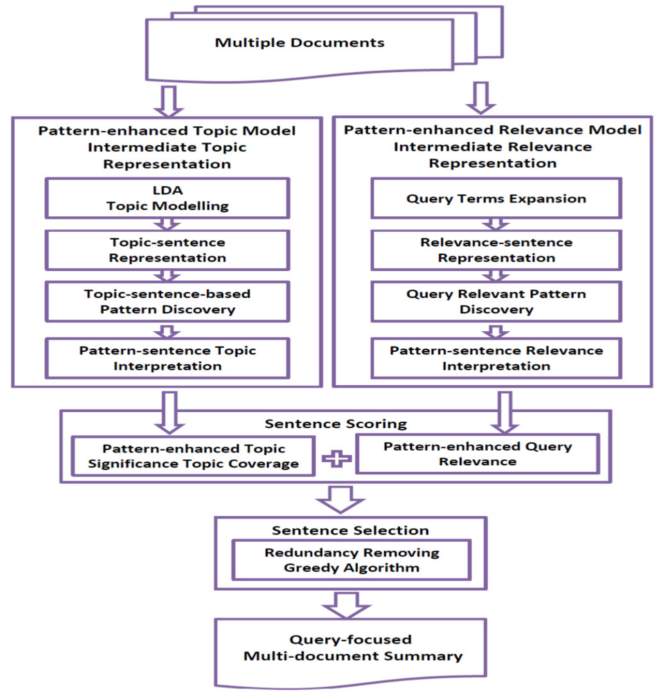



**Fig. 1.** Dual pattern-enhanced representations for query-focused multi-document summarisation (DPRQSum).

### 3.2. Topic representation

To ensure that as many of the main topics in the documents are captured as possible, we utilise a pattern-enhanced topic model [14–16] to represent every individual sentence. When no explicit topic is given, using the advantages of unsupervised LDA topic modelling, we utilise pattern mining as a special query expansion technique to expand implicit topics from no given topics to a set of topical patterns in order to represent the important topics of the given document set. This intermediate topic representation is generated in four steps [15,16].

**Step 1**: Usage of LDA topic modelling[2] to find the topic distribution in a given document collection. The LDA model [46] provides word distribution to represent topics and topic distribution to represent documents. According to the word distribution, each topic is assigned a set of topical words. Table 1 shows examples of word–topic assignments for three topics.

**Table 2**
topic–sentence transactional tables generated from Table 1.

| | $\Gamma_1$ | | $\Gamma_2$ | | $\Gamma_3$ |
|---|---|---|---|---|---|
| $s_{11}$ | $\{w_4, w_{10}, w_{15}\}$ | $s_{12}$ | $\{w_1, w_2, w_5, w_6\}$ | $s_{13}$ | $\{w_1, w_2, w_3, w_6, w_7\}$ |
| $s_{21}$ | $\{w_4, w_5, w_7, w_8\}$ | $s_{22}$ | $\{w_2, w_9, w_{12}\}$ | $s_{23}$ | $\{w_5, w_7, w_{10}\}$ |
| $s_{31}$ | $\{w_8, w_{10}\}$ | $s_{32}$ | $\{w_1, w_{13}, w_{15}\}$ | $s_{33}$ | $\{w_1, w_5, w_6, w_{11}, w_{12}\}$ |
| $s_{41}$ | $\{w_5, w_{10}, w_{11}\}$ | $s_{42}$ | $\{w_5, w_6, w_{12}\}$ | $s_{43}$ | $\{w_1, w_6, w_{11}, w_{14}\}$ |

**Step 2**: Construction of a topic–sentence transactional table for every identified topic in collection $D$. The rows are sentences and the columns are a set of terms assigned to this specified topic by LDA. The transactional table for topic $Z_j$ denoted by $\Gamma_j$ consists of all the sentences that contain any word assigned to topic $Z_j$. Generated from the example word–topic assignments in Table 1, three topic–sentence transactional tables are illustrated in Table 2.

**Step 3**: Discovery of topical patterns. We adopt pattern mining techniques[3] in this step to discover the succinct informative patterns that frequently occur in sentences. DPRQSum utilises frequent closed patterns [49,50] generated from each sentence-based transactional table to represent the identified topic for a given document set. Taking $\Gamma_3$ in Table 2 as an example of the topic–sentence transactional table for topic $Z_3$, with a minimal support threshold $\sigma = 0.5$, all the frequent patterns and closed patterns generated from $\Gamma_3$ are shown in Table 3.

**Definition 1** (*Support*).: The support of a pattern $X$, denoted by $supp(X)$, is the percentage of the transactions in the given transactional table $\Gamma$ that contain $X$.

**Definition 2** (*Frequent Pattern*).: A pattern $X$ is called a frequent pattern if its support $supp(X) >= \sigma$, a given minimal support.

2 In our model, the LDA model is implemented using the MALLET toolkit [45].

3 In pattern discovery, we implement the a priori algorithm [47] to extract frequent patterns, and then apply Pasquier's CLOSE+ algorithm [48] to generate frequent closed patterns.

**Table 3**
Frequent patterns and closed patterns for topic $Z_3$, $\sigma = 0.5$.

| *supp* | Frequent patterns | Closed patterns |
|---|---|---|
| 3/4 | $\{w_1\}, \{w_6\}, \{w_1, w_6\}$ | $\{w_1, w_6\}$ |
| 2/4 | $\{w_5\}, \{w_7\}, \{w_{11}\}, \{w_1, w_{11}\}, \{w_6, w_{11}\}, \{w_1, w_6, w_{11}\}$ | $\{w_5\}, \{w_7\}, \{w_1, w_6, w_{11}\}$ |
| | | ⇑⇑ ***pattern-enhanced topic representation*** for topic $Z_3$ |

**Definition 3** (*Closed Pattern*).: In a transactional table, a pattern $X$ is a closed pattern if there is not a single pattern $X'$ such that (1) $X \subset X'$, and (2) $supp(X) = supp(X')$.

A closed pattern covers all the information described in its subsets, so that the largest range of the associated terms can be revealed [49,50].

**Step 4**: Creation of a ***pattern-enhanced topic representation*** [15] for each identified topic. For instance, topic $Z_j$ can be represented by a set of all closed patterns generated from topic–sentence transactional table $\Gamma_j$, denoted by $X_{Z_j} = \{x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jm_j}\}$, where $m_j$ is the total number of patterns in $X_{Z_j}$. $X_{Z_j}$, called ***topical relevant patterns*** [15,16], is a set of the most representative patterns for topic $Z_j$. In Table 3, the column headed Closed Patterns is $Z_3$'s pattern-enhanced topic representation, and $\{w_1, w_6\}$, $\{w_5\}$, $\{w_7\}$, $\{w_1, w_6, w_{11}\}$ are $Z_3$'s topical relevant patterns.

It is worth mentioning that, although multiple sets of topical relevant patterns need to be discovered for multiple topics, the time complexity of pattern mining here cannot increase significantly, as the size of a given document set is small (e.g., 25) and the size of each topic–sentence transactional table is also limited by the number of sentences assigned to one specific topic via word–topic assignment [14,15].

### 3.3. Query relevance representation

For relevance modelling, we incorporate the query expansion and pattern mining techniques to generate an intermediate pattern-based query relevance representation for sentences. Query expansion, which reformulates a seed query, is widely adopted to improve the retrieval performance in information retrieval operations. In the context of query-focused MDS, as the given query usually consists of very few terms, it is necessary to use query expansion techniques to obtain a better understanding of the query. In the proposed model, we combine query expansion with pattern mining in an innovative 'win-win' manner, i.e., we apply query expansion techniques to improve relevant pattern discovery and utilise the discovered patterns to improve also query expansion.

For an unsupervised extractive summariser with no training data and external resources, it is a reliable choice to use co-occurring terms within the same sentence as query terms for expanding query relevant terms. Therefore, we define a set of ***query relevant terms*** that compromises the original query terms from the given query and the terms co-occurring with the original query terms within the same sentence. An example of query original terms and their co-occurring terms is shown in Table 4, demonstrating how query relevant terms are expanded.

Based on expanded ***query relevant terms***, the intermediate representation for the query relevance can be generated in three steps.

**Step 1**: We construct two transactional tables for one multi-document collection: one is a query relevant transactional table, denoted by $\Gamma_Q$, where the columns are the set of ***query relevant terms*** and each row is a sentence containing at least one original query term, and the second is the background transactional table, denoted by $\Gamma_B$, in which the rows are those sentences that do

**Table 4**
Example query terms and co-occurring terms.

| Terms: | | |
|---|---|---|
| $w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}, w_{13}, w_{14}, w_{15}$ | | |
| Query terms: $w_1, w_6$ | | |
| sentence terms | | terms co-occurring with query terms |
| $s_1$ | $\underline{w_1}, w_3, \underline{w_6}, w_7$ | $w_3, w_7$ |
| $s_2$ | $\underline{w_1}, w_4, w_9$ | $w_4, w_9$ |
| $s_3$ | $w_4, w_8, w_9, w_{12}$ | |
| $s_4$ | $w_3, w_4, w_{10}, w_{13}, w_{15}$ | |
| $s_5$ | $w_4, \underline{w_6}, w_9$ | $w_4, w_9$ |
| $s_6$ | $\underline{w_1}, w_5, \underline{w_6}, w_{11}, w_{12}$ | $w_5, w_{11}, w_{12}$ |
| $s_7$ | $w_4, w_5, w_7, w_8$ | |
| $s_8$ | $\underline{w_1}, \underline{w_6}, w_{11}, w_{14}$ | $w_{11}, w_{14}$ |
| $s_9$ | $w_2, w_4, w_9, w_{13}$ | |
| Co-occurring terms: $w_3, w_4, w_5, w_7, w_9, w_{11}, w_{12}, w_{14}$ | | |
| ***Query relevant terms:*** | | |
| $w_1, w_3, w_4, w_5, w_6, w_7, w_9, w_{11}, w_{12}, w_{14}$ | | |

**Table 5**
Example of transactional tables generated from Table 4.

| | $\Gamma_Q$ | | $\Gamma_B$ |
|---|---|---|---|
| $s_1$ | $\{\underline{w_1}, w_3, \underline{w_6}, w_7\}$ | $s_3$ | $\{w_4, w_8, w_9, w_{12}\}$ |
| $s_2$ | $\{\underline{w_1}, w_4, w_9\}$ | $s_4$ | $\{w_3, w_4, w_{10}, w_{13}, w_{15}\}$ |
| $s_5$ | $\{w_4, \underline{w_6}, w_9\}$ | $s_7$ | $\{w_4, w_5, w_7, w_8\}$ |
| $s_6$ | $\{\underline{w_1}, w_5, \underline{w_6}, w_{11}, w_{12}\}$ | $s_9$ | $\{w_2, w_4, w_9, w_{13}\}$ |
| $s_8$ | $\{\underline{w_1}, \underline{w_6}, w_{11}, w_{14}\}$ | | |

**Table 6**
Frequent closed patterns generated from $\Gamma_Q$ and $\Gamma_B$ in Table 5 for $\sigma = 0.4$.

| *supp* | Closed patterns from $\Gamma_Q$ | *supp* | Closed Patterns from $\Gamma_B$ |
|---|---|---|---|
| 4/5 | $\{w_1\}, \{w_6\}$ | 4/4 | $\{w_4\}$ |
| 3/5 | $\{w_1, w_6\}$ | 3/4 | |
| 2/5 | $\{w_1, w_6, w_{11}\}, \{w_4, w_9\}$ | 2/4 | $\{w_4, w_8\}, \{w_4, w_9\}, \{w_4, w_{13}\}$ |

not contain any original query terms and the columns are all the terms from those sentences. In summary, if a sentence does not contain any query original term, then we assign the sentence to a background sentence set, $S_B$, which comprises the rows in the transactional table $\Gamma_B$, and otherwise, to a query-relevant sentence set, $S_Q$, which comprises table $\Gamma_Q$. Table 5 shows an example of a query relevant transactional table and background transactional table generated from the example of the sentences list in Table 4.

**Step 2**: On the basis of the two transactional tables generated above ($\Gamma_Q$ and $\Gamma_B$), we apply closed pattern mining again (with the minimum support threshold $\sigma$ set to 2.5% of the size of the transactional table, which is the same as in topical pattern mining) to discover query relevant patterns that frequently occur in query-relevant sentence set $S_Q$ and background patterns that exist in the background sentence set $S_B$. This pattern mining step in fact further improves query expansion by discovering patterns of co-occurrences that are the above-chance (the minimum support threshold) frequent occurrences of the original query terms and other terms in one sentence. Taking the two transactional tables in Table 5 as an example, for a minimal support threshold $\sigma = 0.4$, all the frequent closed patterns generated from $\Gamma_Q$ and $\Gamma_B$ are listed in Table 6.

**Step 3**: After the pattern discovery in Step 3, DPRQSum builds a ***pattern-enhanced relevance representation***, which consists of two pattern sets: ***query relevant patterns*** and ***background patterns***. For example, Table 6 is a representation of the pattern-enhanced relevance of the example set of sentences in Table 4; all the closed patterns in the first column are query relevant patterns and those in the second column are background patterns. For a given document set $D$, query relevance now can be represented by a set of all the closed patterns generated from query relevant transactional table $\Gamma_Q$, denoted by $X_Q = \{x_{r1}, x_{r2}, x_{r3}, \ldots, x_{rm_r}\}$, where $m_r$ is the total number of patterns in $X_Q$. Together with $X_Q$, a set of all the closed patterns is extracted from background transactional table $\Gamma_B$, denoted by $X_B = \{x_{b1}, x_{b2}, x_{b3}, \ldots, x_{bm_b}\}$, where $m_b$ is the total number of patterns in $X_B$. DPRQSum utilises these ***background patterns*** as reference features for re-weighting query relevant patterns in a subsequent step. As discussed in the next section, re-weighting is a common query expansion technique, which we adopt to optimise our novel query relevance model.

## 4. Sentence ranking based on balanced topic coverage and query relevance

Sentence ranking is the most important issue under the extractive summarisation framework [51]. In query-focused summarisation, the sentence ranking is determined by a combination of two factors [9]: query relevance, measuring how relevant a sentence is to the given query, and topic coverage/importance, measuring the importance of a sentence in the context of the given document collection in which it appears. In the DPRQSum model, the topic coverage and importance of each sentence are estimated based on ***topical relevant patterns***, and the query relevance is derived from ***query relevant patterns***.

We note that the topic coverage of a sentence here has different meaning from the topic coverage requirement for query-focused MDS, as the latter is at the summary document level, which requires that the generated summary reserve as many important topics as possible for the given document set, while the former indicates how much information an individual sentence covers, regardless off the number of topics covered by the information. Furthermore, if a sentence presents only one topic but this topic is very important to the entire document set, we still consider this sentence a good candidate for forming the summary. Therefore, in the following ranking steps, we define the ***topical significance*** of a sentence as an evaluation metric that indicates both the amount of important information the sentence contains, the so-called *topic coverage*, and the distinctiveness of the information presented in the sentence, the so-called *topic novelty*.

### 4.1. Sentence topical significance score

Being founded on the generated pattern-enhanced topic representation for each identified topic, DPRQSum measures the importance of every sentence for its associated topics by utilising the topic significance of the matched topical relevant patterns. According to Gao et al. [14], 'the significance of one pattern is determined not only by its statistical significance, but also by its size, since the size of the pattern indicates the specificity level'. They defined the pattern specificity in their model [14] as follows.

**Definition 4** (*Pattern Specificity*)**.** The specificity of a pattern $X$ is defined as a power function of the pattern length with an exponent less than 1, denoted by $spe\,(X)$, $spe\,(X) = a|X|^m$, where $a$ and $m$ are constant real numbers and $0 < m < 1$.

In the proposed DPRQSum model, when the multiple latent topics contained in the given documents set have been identified, the topical significance score of a sentence is determined by the sum of the topic significance of every topic to the sentence.

Let $s$ be a sentence, $Z_j$ be an identified topic of a document set, $x_{jk} \in X_{Z_j}$, $k = 1, \ldots, m_j$ be a topical relevant pattern and $supp(x_{jk})$ be the corresponding supports of $x_{jk}$; then, the topic significance of $Z_j$ to $s$ is defined as

$$T_{sig}\left(Z_j, s\right) = \sum_{k=1}^{m_j} spe\left(x_{jk}\right) \times supp(x_{jk}) \times \tau\left(x_{jk}, s\right) = \sum_{k=1}^{m_j} \left|x_{jk}\right|^{0.5} \times supp(x_{jk}) \times \tau\left(x_{jk}, s\right) \quad (1)$$

where $\tau\left(x_{jk}, s\right) = 1$ if $x_{jk} \in s$; otherwise, $\tau\left(x_{jk}, s\right) = 0$, and we set the scale of pattern specificity to 0.5 as in [14].

By means of the above function, sentences are represented as a vector of topic significances over a set of topics, which makes it feasible to analyse topic coverage and novelty in the topic space. Specifically, given a sentence $s$ with $V$ identified topics, DPRQSum computes the ***topical significance*** indicating both topic coverage and novelty as

$$T_{sig}\,(s) = \sum_{j=1}^{V} T_{sig}\left(Z_j, s\right) \times log_2(1 + \frac{V}{v_s}) \quad (2)$$

where $V$ is the total number of identified topics and $v_s$ is the number of the sentence's associated topics, of which at least one of the topical relevant patterns appears in this sentence (i.e., $T_{sig}\left(Z_j, s\right) > 0$). The part $log_2(1 + \frac{V}{v_s})$ in fact is the ***inverse topic frequency***, similar to the IDF in Term Frequency * Inverse Document Frequency (TF*IDF), introduced here as 'topic specificity' to balance topic coverage and novelty for a given sentence. By combining this topic specificity with topical patterns coverage, a sentence covering more topical patterns from fewer topics will be given a higher score for its topical significance.

### 4.2. Sentence query relevance score

To measure the relevance of a sentence to a given query, based on our novel pattern-enhanced query relevance representation, DPRQSum utilises the significance of the matched query relevant patterns in a given sentence. The proposed model involves a **re-weighting** process for every query relevant pattern, which is performed before the calculation of a sentence's query relevance. This re-weighting process borrows the theory of three-way decisions [52–54] to treat query relevant patterns as an approximate concept (a rough set). We use a pair of lower and upper approximations to divide query relevant patterns into three disjoint regions: the patterns in the lower approximation, defined as **true-relevant patterns**; the patterns in the boundary between the upper and lower approximations, defined as **vaguely-relevant patterns**; and the patterns in the upper approximation, defined as **false-relevant patterns**.

To apply the rough set theory-based three-way decision rules, the crucial problem is how to estimate the value of the odds ratio of the relevance. Li et al. [52,55] simplified an odds function and the decision rules for information retrieval as

$$Odds\,(x_r) = \frac{P\,(x_r)}{1 - P\,(x_r)}, \quad (3)$$

where $P\,(x_r)$ is the probability that the given element is relevant and $1 - P\,(x_r)$ is the probability that the element is not relevant. Then, with the two decision parameters $\gamma_1$ and $\gamma_2$, three decision rules are defined

$$\begin{aligned} Odds\,(x_r) > \gamma_1 &\mapsto PositiveRegion \\ Odds\,(x_r) < \gamma_2 &\mapsto NegativeRegion \\ \gamma_2 < Odds\,(x_r) < \gamma_1 &\mapsto BoundaryRegion \end{aligned} \quad (4)$$

As a probabilistic method, the odds function in the above is estimated based on a simplest independence assumption, i.e., the distribution of elements is independent in both the relevant set and all sets, and also it chooses to obey a relevance ordering principle: the probable relevance is based only on the presence of the search elements in the given set [56].

However, this odds function ignores the effect of the independence of distribution in a non-relevant set. It also has been proved that the probable relevance should be based on both the presence and absence of elements in a given set [56]. Therefore, in our model, we propose a BM25-like re-weighting algorithm to adjust the representative power, i.e., the odds of query relevance for each query relevant pattern. Given a set of query relevant patterns $X_Q$ and a set of background patterns $X_B$ presented in Section 3.3, the BM25-like equation is defined as

$$Odds\left(x_r\right) = log\left(\frac{\frac{supp_Q(x_r)+0.5}{1-supp_Q(x_r)+0.5}}{\frac{supp_B(x_r)+0.5}{1-supp_B(x_r)+0.5}}\right), \tag{5}$$

where for all $x_r \in X_Q$, $supp_Q(x_r)$ is the percentage support in $X_Q$ and $supp_B(x_r)$ is the corresponding percentage support in background patterns $X_B$ if $x_r \in X_B$ as well, and otherwise $supp_B(x_r) = 0$. The constant number 0.5 is used to avoid zero-division-error.

It is easy to see that the value of $Odds\left(x_r\right)$ is in the range $(-1, 1)$. Therefore, we make further simplifications here by setting two decision parameters, called the two thresholds of relevance, $0 < \gamma_1 < 1$ and $-1 < \gamma_2 < 0$. Then, we can define the following three regions of query relevant patterns:

- **True-relevant patterns**: a query relevant pattern $x_r \in X_Q$ is a true-relevant pattern if $Odds\left(x_r\right) > \gamma_1$, as it appears considerably more frequently in query relevant sentences than in background sentences.
- **Vaguely-relevant patterns**: a query relevant pattern $x_r \in X_Q$ is a vaguely-relevant pattern if $\gamma_2 \le Odds\left(x_r\right) \le \gamma_1$, as it appears too frequently in both query relevant sentences and background sentences.
- **False-relevant patterns**: a query relevant pattern $x_r \in X_Q$ is a false-relevant pattern if $Odds\left(x_r\right) < \gamma_2$, as it appears more frequently in background sentences than in query relevant sentences.

To maximise the representative power of query relevant patterns, we introduce a **'discriminating' strategy** in the re-weighting process of the query relevant pattern. The basic idea of the strategy is to preserve true-relevant patterns, ignore vaguely-relevant patterns by assigning zero value to their weight, and punish false-relevant patterns by using their negative weights in query relevance ranking for each sentence.

After the above re-weighting process including the odds function, decision rules and the discriminating strategy have been applied, the final re-weighted weight of each query relevant pattern $W\left(x_r\right)$ also falls into the three regions of positive, negative and zero.

Algorithm 1 summarises the discussed re-weighting process for query relevant patterns. The time complexity of the re-weighting algorithm is linear to the size of the query relevant patterns set, i.e., $O(m_r)$, where $m_r$ is the total number of patterns in $X_Q$ (see Section 3.3).

**Algorithm 1:** DPRQSum query relevant pattern re-weighting

```
Input : The set of query relevant patterns X_Q; the set of
        background patterns X_B; two thresholds of relevance
        γ1 and γ2, and 0 < γ1 < 1, −1 < γ2 < 0;
Output: The set of re-weighted query relevant patterns X'_Q
        with re-weighted weight of each relevant pattern
        W (x_r)
; X'_Q := ∅
for each x_r ∈ X_Q do
    Odds (x_r) := Eq. (5)
    if γ2 ≤ Odds (x_r) ≤ γ1 then
        W (x_r) := 0 //zero weight, no need to put in X'_Q
    end
    else
        // Odds (x_r) > γ1 > 0 or Odds (x_r) < γ2 < 0
        W (x_r) := Odds (x_r)
        X'_Q := X'_Q ∪ {x_r}
    end
end
return X'_Q
```

In DPRQSum, the query relevance of sentences is determined not only by the coverage of query relevant patterns with the updated pattern weight, but also by the number of original query terms contained in the given sentence. Therefore, we estimate the query relevance of sentences as

$$\begin{aligned} Q_{rel}\left(s\right) &= \sum_{r=1}^{m_r} spe\left(x_r\right) \times W\left(x_r\right) \times \tau\left(x_r, s\right) \times t_q\left(s\right)^{0.5} \\ &= \sum_{r=1}^{m_r} \left|x_r\right|^{0.5} \times W\left(x_r\right) \times \tau\left(x_r, s\right) \times t_q\left(s\right)^{0.5} \end{aligned} \tag{6}$$

where $t_q\left(s\right)$ denotes the number of original query terms contained in $s$ and $\tau\left(x_r, s\right) = 1$ if $x_r \in s$, and otherwise, $\tau\left(x_r, s\right) = 0$.

### 4.3. Sentence ranking score

For final sentence ranking, we apply a normalised weighted linear combination of the topical significance score (Eq. (2)) and the query relevance score (Eq. (6)), with two additional features, sentence position and sentence length, also integrated. We utilise the sentence position and length features to bias sentence selection towards the first sentence in the document and sentences with a length close to average respectively. Thus, we update the score of each sentence $s$ using

$$Score\left(s\right) = \left(1 - \rho - \iota\right) \times Rank\left(s\right) + \rho \times POS\left(s\right) + \iota \times SL\left(s\right) \tag{7}$$

where $Rank\left(s\right)$ is the normalised (using the maximal value of the topical significance scores and query relevance scores of all sentences respectively to scale scores to between 0 and 1) weighted linear combination of the topical significance score (Eq. (2)) and the query relevance score (Eq. (6)), defined as

$$Rank\left(s\right) = \left(1 - \varrho\right) \times T_{sig}\left(s\right) + \varrho \times Q_{rel}\left(s\right) \tag{8}$$

where $\varrho$ denotes the weight coefficient of the query relevance model. $POS\left(s\right)$ is the position importance of sentence $s$:

$$POS\left(s\right) = \begin{cases} 1 & \text{if } s \text{ is the first sentence in a document} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and $SL\left(s\right)$ is the length weight of sentence $s$, estimated by

$$SL\left(s\right) = 1 - \frac{\left|l\left(s\right) - L_{avg}\right|}{L_{avg}}, \tag{10}$$

where

$$L_{avg} = \frac{\sum_{s_i \in D} l(s_i)}{|D|} \tag{11}$$

and $\rho$ and $\iota$ are constant values (parameters) to indicate the weight coefficient of position importance $POS(s)$ and length weight $SL(s)$ respectively. (In this study, we set $\rho = 0.06$; $\iota = 0.04$).

The summarised sentence ranking process is depicted in Algorithm 2. With the input of the number of identified topics $V$ and the number of sentences $I$, the sentence ranking algorithm has complexity $O(V * I)$.

**Algorithm 2:** DPRQSum sentence ranking

**Input** : The document set $D = \{s_1, s_2, \ldots s_i\}$; $V$ identified topics, the set of $V$ set of topical relevant patterns $X_Z = \{X_{Z_1}, X_{Z_2}, \ldots, X_{Z_V}\}$; the set of re-weighted query relevant patterns $X'_Q$; sentence pool size $n$, and parameters $\rho, \iota, \varrho$;

**Output**: The set of top-$n$ sentence-*Rank* pairs $SR$;

$SR := \emptyset$
**for** *each* $s \in D$ **do**
  $T_{sig}(s) := 0$
  $v_s := 0$
  **for** *each* $Z_j \in V$ **do**
    $T_{sig}(Z_j, s) =$ Eq. (1)
    **if** $T_{sig}(Z_j, s) > 0$ **then**
      $T_{sig}(s) := T_{sig}(s) + T_{sig}(Z_j, s)$
      $v_s := v_s + 1$
    **end**
  **end**
  $T_{sig}(s) := T_{sig}(s) \times log_2(1 + \frac{V}{v_s})$ // see Eq. (2)
  $Q_{rel}(s) :=$Eq. (6)
**end**
$max\_T_{sig} := max(T_{sig}(s) \,|\, s \in D)$
$max\_Q_{rel} := max(Q_{rel}(s) \,|\, s \in D)$
**for** *each* $s \in D$ **do**
  $Rank(s) = (1 - \varrho) \times (T_{sig}(s) / max\_T_{sig}) + \varrho \times (Q_{rel}(s) / max\_Q_{rel})$ // see Eq. (8)
**end**
Sort the sentences $\{s_1, s_2, \ldots s_i\}$ by their final ranking score $Score(s)$
Add top-$n$ sentence–Rank pairs $\{s : Score(s)\}$ to $SR$
**return** $SR$

### 4.4. Greedy sentence selection with redundancy reduction

For MDS tasks, where the maximal information coverage and relevance are preserved, redundancy removal becomes a fateful process in the sentence selection stage. For sentence selection, the proposed DPRQSum adopts the greedy approach of maximal marginal relevance (MMR) [57] and a diversity penalty imposition [58,59] is applied. The sentence having the highest ranking score in the sentences pool is selected first into the final summary. Then, using the following equation as formulated in SRRank [59], the sentences with the highest diversity penalty-imposed score in the remaining sentences are added one by one to the summary.

$$Score^{(t+1)}(s_i) = Score^{(t)}(s_i) - \delta \times CosSim(s_i, s_j) \times Score_{SR}(s_j) \tag{12}$$

where $s_j$ is the last selected sentence with the highest overall score, and then $s_i$ is the remaining sentence that will be penalised, and $CosSim(s_i, s_j)$ is the normalised cosine similarity between the two sentences $s_i$ and $s_j$. The penalty weight parameter $\delta$ here can be empirically tuned for the best result. The selection and penalisation are performed iteratively until the summary length reaches the limit [15].

**Table 7**
Description of the DUC datasets for query-focused multi-document summarisation.

| | DUC 2006 | DUC 2007 |
|---|---|---|
| Number of topics | 50 | 45 |
| Document number per topic | 25 | 25 |
| Avg. sentence number per topic | 705 | 499 |
| Data source | AQUAINT | AQUAINT |
| Summary length | 250 words | 250 words |

## 5. Experiment methodology

In this section, we describe the setup of three sets of experiments to verify the following three hypotheses:

- *Hypothesis H1.* The proposed DPRQSum model is effective for query-focused MDS as compared with other state-of-the-art unsupervised extractive summarisation systems.
- *Hypothesis H2.* The dual pattern-enhanced representations for topic importance and query relevance perform better than the sole pattern-based representation model for query-focused MDS task.
- *Hypothesis H3.* The use of re-weighted query relevant patterns can represent sentences' query relevance more accurately than using query relevant patterns alone.

### 5.1. Datasets

For the evaluation, we used two datasets of DUC 2006 and DUC 2007, both of which are open benchmark datasets for automatic summarisation evaluation. Query-focused MDS is the main task of DUC 2006 and DUC 2007, and models real-world complex question answering in which a question cannot be answered by simply stating a name, date, quantity, etc. The DUC 2006 and DUC 2007 datasets consist of 50 and 45 topics respectively, and each topic includes a document collection of 25 news articles and a short description of a topical query. The task is to create a summary containing no more than 250 words for each document set to answer the question(s) in the topic statement. There are at least four human written reference summaries provided in each document collection for evaluation. The datasets are briefly described in Table 7.

### 5.2. Evaluation metrics

We used the recall-oriented understudy of gisting evaluation (ROUGE)[4] toolkit [60] to evaluate the performance of the proposed summarisation method. The DUC adopted ROUGE as the official evaluation metric for automatic document summarisation. ROUGE metrics automatically determine the quality of a summary by counting the number of overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and an 'ideal' summary or a set of 'ideal' summaries created by humans. In the experiment, we recorded[5] the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores with their corresponding 95% confidential intervals. ROUGE-1 and ROUGE-2 measure the overlap of unigram and bigrams to evaluate the extent to which the testing summary is consistent with human judgements [61] and ROUGE-SU4 examines the overlap of skip-bigrams with a maximum skip distance of four. Skip-bigrams measure the overlap of any pair words in their sentence order allowing arbitrary gaps between them. In general, the higher the ROUGE scores, the more similar the summaries are to 'ideal' summaries, and therefore the better they are.

4 http://www.berouge.com/Pages/default.aspx

5 The command we used was: ROUGE-1.5.5.pl -e data -l 250 -n 4 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d.

### 5.3. Baselines

To evaluate the overall performance of the proposed DPRQSum model, we compared it with that of a variety of baseline methods, including NIST baselines, and several state-of-the-art unsupervised extractive query-focused summarisation systems, which are typically based on different intermediate representations, such as topic model, graph model and group sparse learning model.

- **NIST baseline — Lead** [62] is an unsupervised extraction-based summarisation method, taking the leading sentences one by one (up to 250 words) in the chronologically ordered documents for each document collection. It is a standard baseline provided by NIST in both DUC 2006 and DUC 2007.
- **DUC 2007 baseline — clustering, linguistics, and statistics for summarisation yield (CLASSY04)** [63] consists of two core components, a hidden Markov model to select sentences from each document and a pivoted QR algorithm to generate a multi-document summary. The CLASSY summarisation system was the previous best system in terms of the ROUGE-1 metric for the DUC 2004 main task, and then it became a high-performance baseline in the DUC 2007 main task.
- **HIERSUM** [32] is an unsupervised extractive summarisation method that uses a hierarchical LDA-style model to represent content specificity as a hierarchy of topic vocabulary distribution.
- **DsR-Q** [64] is an unsupervised document-sensitive graph model that exploits document–document and document–sentence relations to distinguish intra-document sentence relations from inter-document sentence relations and considers the influence of the entire document sets on the individual sentence evaluations.
- **SpOpt-$\Delta$** [65] is an unsupervised method for MDS via sparse optimisation with a decomposable convex objective function that introduces a sentence dissimilarity term to encourage diversity. Although it has a variant that utilises sentence compression, which shows better results, we compared our model with only the base **SpOpt-$\Delta$** method, since in our model a similar sentence compression step is not applied.
- **SGS** [66] is an unsupervised MDS system with a group sparse learning framework that maximally reconstructs the original documents by minimising the approximation error and simultaneously selects summary sentences with the learnt group structure information among sentences.

As our model is unsupervised, we did not compare it with supervised methods, except two NIST reported baselines. In order to provide an overall picture, we also present the results of the average of the given human-created summaries (denoted by **avgHuman**), the average of the participant systems (denoted by **avgSys**) and the best of the participant systems (denoted by **bestSys**) in the DUC 2006 and 2007 main tasks competition.

## 6. Evaluation results

### 6.1. Overall performance

We first conducted experiments to verify the effectiveness of the proposed dual pattern-enhanced representations model and ranking algorithm for query-focused MDS tasks (to examine H1). All the parameters in the proposed model are empirically tuned for the best result. For the compared approaches, we list the best results reported in the original literature. The overall performance comparisons on the DUC 2006 and DUC 2007 datasets are shown in Table 8.

As seen in the table, in terms of all three measures the proposed method (DPRQSum) performed considerably better than avgSys (out of 32 in DUC 2006 and 2007) and the DUC baselines, including the simple baseline Lead model on both datasets and the high-performance generic baseline CLASSY04 on DUC 2007, which demonstrates that our proposed method is effective for query-focused MDS. Although the results of the proposed model are not better than those of bestSys, it is worth noting that our model needs no training examples or external resources, whereas bestSys, IIIT Hyderabad [67,68], integrated large scale Web resources and some hand-crafted steps in the processing.

We also performed a paired $t$-test [69] between the ROUGE scores of our model and those of the NIST baselines on both the DUC 2006 and 2007 datasets to examine the significance of the difference between the proposed model and the NIST baselines. In Table 9, the associated $P$ values are listed for DPRQSum.[6] As seen in the table, the proposed pattern-based model significantly outperforms the NIST baseline models, as all the $P$-values of the paired $t$-test are considerably less than 0.05 (5% significance level) in terms of all three ROUGE metrics on both the DUC 2006 and DUC 2007 datasets.

To further prove the overall performance of our unsupervised pattern-based approach, we compared DPRQSum with several state-of-the-art unsupervised approaches including a topic model-based approach HIERSUM, a graph model DsR-Q, a sparse optimisation model SpOpt-$\Delta$ and a group sparse learning framework SGS. We were pleased to observe that the proposed method is superior to these state-of-the-art unsupervised methods on both datasets. Among the state-of-the-art methods, HIERSUM is the strongest baseline, performing better than the other methods. The proposed model outperforms HIERSUM in terms of five out of six metrics on the two datasets, which demonstrates that the proposed DPRQSum achieved an outstanding performance.

### 6.2. Effectiveness of dual pattern-enhanced representations

The objective of the second set of experiments was to examine the individual or combined contributions of the two pattern-based components. **DPRQSum-T** and **DPRQSum-Q** are two variants of the proposed model DPRQSum. The former uses the *topical relevant patterns* only (without query relevant patterns and query relevance scores in sentence ranking) and the latter uses the *query relevant patterns* only (without topical relevant patterns and topical significance scores in sentence ranking). Table 10 shows the results of the average recall scores of the three ROUGE metrics on the DUC 2006 and 2007 datasets.

As shown in Table 10, the performance of all three proposed pattern-enhanced models is considerably better than that of aveSys and the NIST baselines on both the DUC 2006 and DUC 2007 datasets. This demonstrates the effectiveness of the proposed pattern-enhanced representation for query-focused MDS tasks. The results also show that the query relevant patterns are more beneficial than topical relevant patterns for the query-focused summarisation. However, the best results are achieved when both topical patterns and query relevant patterns are employed in DPRQSum with a 1.01–6.12% improvement in the ROUGE performance over the two single pattern-based representation models. These noteworthy results verify that dual pattern-enhanced representations, not only for topical significance but also for query relevance, can improve the performance for query-focused MDS.

[6] We could not perform the $t$-test between the ROUGE results of the proposed model and those of the other comparison state-of-the-art approaches listed above, because the detailed ROUGE results of these approaches are not available in their published reports.

**Table 8**
Experimental results of ROUGE evaluation on DUC 2006 and 2007 datasets.

| System | Recall | | |
|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-SU4 |
| **DUC 2006** | | | |
| **avgHumen** | 0.45771 | 0.11250 | 0.17060 |
| **avgSys** | 0.37256 | 0.07380 | 0.12928 |
| **bestSys** | 0.41108 [0.40488-0.41706] | 0.09558 [0.09144-0.09977] | 0.15529 [0.15126-0.15906] |
| **NIST Baseline — Lead** | 0.30217 [0.29229-0.31193] | 0.04926 [0.04560-0.05375] | 0.09788 [0.09339-0.10215] |
| **HIERSUM** | 0.401 | 0.086 | 0.143 |
| **DsR-Q** | 0.39550 [0.38970-0.40120] | 0.08990 [0.08570-0.09430] | 0.14270 [0.13910-0.14640] |
| **SpOpt-$\Delta$** | 0.39962 | 0.08682 | 0.14227 |
| **SGS** | 0.38798 | 0.08361 | 0.13882 |
| **DPRQSum** | **0.40551** [0.39450-0.41776] | **0.09228** [0.08381-0.10111] | **0.14966** [0.14223-0.15773] |
| **DUC 2007** | | | |
| **avgHumen** | 0.47948 | 0.14099 | 0.19158 |
| **avgSys** | 0.39728 | 0.09486 | 0.14747 |
| **bestSys** | 0.44508 [0.43793-0.45211] | 0.12448 [0.11961-0.12925] | 0.17711 [0.17244-0.18176] |
| **NIST Baseline — Lead** | 0.31250 [0.30328-0.32176] | 0.06039 [0.05636-0.06435] | 0.10507 [0.10076-0.10919] |
| **CLASSY04** | 0.40562 [0.39871-0.41250] | 0.09382 [0.08927-0.09805] | 0.14641 [0.14224-0.15066] |
| **HIERSUM** | 0.424 | **0.118** | 0.167 |
| **DsR-Q** | 0.42190 [0.41420-0.42940] | 0.11230 [0.10730-0.11710] | 0.16820 [0.15980-0.16990] |
| **SpOpt-$\Delta$** | 0.42360 | 0.11109 | 0.16474 |
| **SGS** | 0.40991 | 0.10264 | 0.15639 |
| **DPRQSum** | **0.43404** [0.42251-0.44514] | 0.11683 [ 0.10836-0.12544] | **0.17026** [0.16284-0.17741] |

**Table 9**
$P$-values of paired $t$-test on DUC 2006 and DUC 2007 as compared with the NIST baselines ($alpha = 0.05$) - DPRQSum.

| | DUC 2006 | 50 clusters | | DUC 2007 | 45 clusters | |
|---|---|---|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-SU4 | ROUGE-1 | ROUGE-2 | ROUGE-SU4 |
| **NIST Baseline — Lead** | | | | | | |
| **one-tail** | 5.96E−11 | 1.32E−09 | 2.51E−11 | 4.20E−17 | 9.63E−18 | 7.07E−20 |
| **two-tail** | 1.19E−10 | 2.64E−09 | 5.01E−11 | 8.39E−17 | 1.93E−17 | 1.41E−19 |
| **NIST Baseline — CLASSY04** | | | | | | |
| **one-tail** | | | | 6.00E−05 | 3.67E−06 | 2.20E−07 |
| **two-tail** | | | | 0.000120 | 7.35E−06 | 4.41E−07 |

**Table 10**
Comparison with singleton pattern-based representation models.

| System | Recall | | |
|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-SU4 |
| **DUC 2006** | | | |
| **avgSys** | 0.37256 | 0.07380 | 0.12928 |
| **NIST Baseline — Lead** | 0.30217 [0.29229-0.31193] | 0.04926 [0.04560-0.05375] | 0.09788 [0.09339-0.10215] |
| **DPRQSum-T** | 0.39389 [0.38271-0.40535] | 0.08561 [0.07743-0.09466] | 0.14103 [0.13369-0.14902] |
| **DPRQSum-Q** | 0.40266 [0.39177-0.41415] | 0.09186 [0.08302-0.10063] | 0.14816 [0.14105-0.15560] |
| **DPRQSum** | 0.40551 [0.39450-0.41776] | 0.09228 [0.08381-0.10111] | 0.14966 [0.14223-0.15773] |
| **DUC 2007** | | | |
| **avgSys** | 0.39728 | 0.09486 | 0.14747 |
| **NIST Baseline — Lead** | 0.31250 [0.30328-0.32176] | 0.06039 [0.05636-0.06435] | 0.10507 [0.10076-0.10919] |
| **DPRQSum-T** | 0.41927 [0.40648-0.43223] | 0.10622 [0.09797-0.11486] | 0.16077 [0.15303-0.16913] |
| **DPRQSum-Q** | 0.42561 [0.41477-0.43655] | 0.11045 [0.10264-0.11941] | 0.16407 [0.15707-0.17131] |
| **DPRQSum** | 0.43404 [0.42251-0.44514] | 0.11683 [ 0.10836-0.12544] | 0.17026 [0.16284-0.17741] |

### 6.3. *Effectiveness of re-weighting*

As an additional contribution of this paper is the novel re-weighting algorithm for query relevance representation, we conducted a set of experiments to evaluate the effectiveness of our innovative re-weighting algorithm in query relevance ranking. We built two variants of DPRQSum model. One, denoted by **DPRQSum-no/rw**, replaces re-weighted query-relevant-patterns with original query-relevant-patterns in query relevance ranking, and the second, denoted by **DPRQSum-no/grp**, eliminates the 'discriminating' strategy in the re-weighting process by discarding the two thresholds of relevance $\gamma_1$ and $\gamma_2$ in Algorithm 1. The comparison results on the two DUC datasets are shown in Table 11.

As shown in the table, even when the entire re-weighting process for query relevant patterns is eliminated, the two variants of the proposed dual pattern-enhanced model can significantly outperform aveSys and the NIST baseline for all ROUGE metrics on the DUC 2006 and DUC 2007 evaluation. While DPRQSum-no/grp performs a bit better than DPRQSum-no/rw, DPRQSum is superior to DPRQSum-no/grp with an improvements of 2.02–3.98% in terms of ROUGE-2 and 1.35–2.39% in terms of ROUGE-SU4. These promising improvements verify hypothesis H3, i.e., the novel re-weighting process for query relevant patterns can render query relevance ranking more accurate.

To show that the effect of the proposed re-weighting algorithm is significant, Table 12 presents the $P$-values produced by the paired $t$-test between the ROUGE scores of DPRQSum and those of the two variants, DPRQSum-no/rw and DPRQSum-no/grp, which do not apply re-weighting and/or a 'discriminating' strategy. It is clearly seen in the table that the $P$-values of the paired $t$-test for the comparison of DPRQSum and DPRQSum-no/rw on both datasets

**Table 11**
Comparison with query relevance ranking methods without re-weighting.

| System | Recall | | |
|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-SU4 |
| **DUC 2006** | | | |
| **avgSys** | 0.37256 | 0.07380 | 0.12928 |
| **NIST Baseline — Lead** | 0.30217 [0.29229-0.31193] | 0.04926 [0.04560-0.05375] | 0.09788 [0.09339-0.10215] |
| **DPRQSum-no/rw** | 0.39893 [0.38717-0.40993] | 0.08861 [0.08124-0.09654] | 0.14525 [0.13800-0.15254] |
| **DPRQSum-no/grp** | 0.40245 [0.39081-0.41430] | 0.09045 [0.08268-0.09882] | 0.14767 [0.14038-0.15529] |
| **DPRQSum** | 0.40551 [0.39450-0.41776] | 0.09228 [0.08381-0.10111] | 0.14966 [0.14223-0.15773] |
| **DUC 2007** | | | |
| **avgSys** | 0.39728 | 0.09486 | 0.14747 |
| **NIST Baseline — Lead** | 0.31250 [0.30328-0.32176] | 0.06039 [0.05636-0.06435] | 0.10507 [0.10076-0.10919] |
| **DPRQSum-no/rw** | 0.42698 [0.41513-0.43872] | 0.11228 [0.10440-0.12104] | 0.16617 [0.15863-0.17394] |
| **DPRQSum-no/grp** | 0.42740 [0.41524-0.43929] | 0.11236 [0.10435-0.12118] | 0.16628 [0.15868-0.17398] |
| **DPRQSum** | 0.43404 [0.42251-0.44514] | 0.11683 [ 0.10836-0.12544] | 0.17026 [0.16284-0.17741] |

**Table 12**
*P*-values of paired *t*-test on DUC 2006 and DUC 2007 as compared with DPRQSum-no/rw and DPRQSum-no/grp ($alpha = 0.05$).

| | DUC 2006 | 50 clusters | | DUC 2007 | 45 clusters | |
|---|---|---|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-SU4 | ROUGE-1 | ROUGE-2 | ROUGE-SU4 |
| **DPRQSum-no/rw** | | | | | | |
| **one-tail** | 0.00836222 | 0.01930339 | 0.00427133 | 0.00071826 | 0.00394110 | 0.00383250 |
| **two-tail** | 0.01672445 | 0.03860678 | 0.00854267 | 0.00143653 | 0.00788220 | 0.00766501 |
| **DPRQSum-no/grp** | | | | | | |
| **one-tail** | 0.05966885 | 0.11406360 | 0.05324117 | 0.00245042 | 0.00298424 | 0.00440282 |
| **two-tail** | 0.11933770 | 0.22812720 | 0.10648234 | 0.00490085 | 0.00596848 | 0.00880565 |

are considerably less than 0.05 in terms of the ROUGE-1, ROUGE-2 and ROUGE-SU4 metrics. This is strong evidence against the null hypothesis, indicating that the re-weighting process applied in DPRQSum has a statistically significant effect on the performance of query-focused MDS tasks. However, for the comparison of DPRQSum and DPRQSum-no/grp, which applies the re-weighting process partially without implementing the 'discriminating' strategy, the *P*-values of the paired *t*-test on the DUC 2006 dataset are greater than 0.05, whereas the *P*-values on the DUC 2007 dataset are less than 0.05. Based on these results, although the impact of the 'discriminating' strategy is not very significant, as the whole re-weighting process is applied, it still shows a promising improvement in the overall performance for query-focused MDS tasks.

## 6.4. Parameter tuning

### 6.4.1. Weight coefficient of the query relevance model

In the proposed model, $\varrho$ (in Eq. (8)) is used to tune the trade-off between topical significance and query relevance. We conducted systematic experiments to determine the appropriate values for $\varrho$ that yield the best performance of the system. We varied the weight $\varrho$ from 0 to 1. Fig. 2 shows the ROUGE-2 and ROUGE-SU4 curves for the DUC 2006 and DUC 2007 datasets respectively. In general, with an increase in $\varrho$, the performance first increases, reaches its peak value and then is degraded. It is of interest that the best values of $\varrho$ differ for DUC 2006 and DUC 2007. For the DUC 2006 dataset, a value of $\varrho$ around 0.8 can attain the best ROUGE-2 and ROUGE-SU4 results, while for DUC 2007, $\varrho = 0.5$ produces peak values of ROUGE scores. These observations correspond to the fact that the query relevance importance differs for different datasets; i.e., the extraction of the same length of summaries from more sentences (e.g., 250 words out of 700 sentences in DUC 2006) needs stricter rules to be applied on the query relevance than the extraction from fewer sentences (e.g., 250 words out of 500 sentences in DUC 2007; see Table 7).

### 6.4.2. Thresholds of relevance $\gamma_1$ and $\gamma_2$

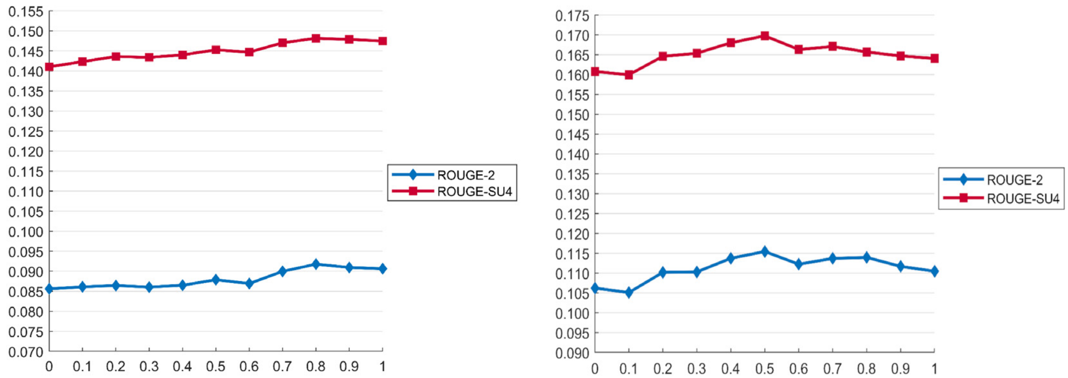
$\gamma_1$ and $\gamma_2$ are two thresholds of relevance, which are employed to group query relevant patterns in the re-weighting process in order to apply the novel 'discriminating' strategy (see Algorithm 1). We conducted a set of experiments with different $\gamma_1$ and $\gamma_2$ combinations to evaluate their influence. The combination of the two thresholds makes it difficult to find a global optimised solution. Therefore, a gradient search strategy was used. We first set negative threshold $\gamma_2$ ($-1 < \gamma_2 < 0$) to double the average weight of all the query relevant patterns that have negative weight (denoted as $n_{avg}$). Then, the performance was evaluated using different values of $\gamma_1$ ($0 < \gamma_1 < 1$) derived from varying ratios of the average weight of all the positive query relevant patterns (denoted by $p_{avg}$), where the varying ratios were in the range 0.1–2.0.

Figs. 3 and 4 present the ROUGE-2 and ROUGE-SU4 evaluation results of DPRQSum on DUC 2006 and DUC 2007 respectively for different values of $\gamma_1$, which were assigned by varying the ratios of $p_{avg}$, and Figs. 5 and 6 show the ROUGE scores with respect to different values of $\gamma_2$ that were derived from varying the ratios of $n_{avg}$ with $\gamma_1$ set to the best value.

As can be seen, the curves of ROUGE-2 and ROUGE-SU4 for the positive threshold $\gamma_1$ on both datasets are in a wider range (a 0.00394 difference for ROUGE-2 and 0.00452 difference for ROUGE-SU4 on DUC 2006, and a 0.00652 divergence of ROUGE-2 and 0.00596 divergence of ROUGE-SU4 on DUC 2007) than for the negative threshold $\gamma_2$ (a 0.00036 difference of ROUGE-2 and 0.00021 difference for ROUGE-SU4 on DUC 2006, and a 0.00064 divergence for ROUGE-2 and 0.00045 divergence for ROUGE-SU4 on DUC 2007). This indicates that the performance effect of the positive threshold $\gamma_1$ seems stronger than that of the negative threshold $\gamma_2$.

## 7. Conclusions

In this paper, we proposed a novel query-focused MDS approach *DPRQSum* that employs dual pattern-enhanced models to represent topical significance and query relevance for documents and sentences. In addition to the pattern-based topic model, which




(a) DUC 2006 dataset

(b) DUC 2007 dataset

**Fig. 2.** Performance with respect to the weight coefficient of the query relevance model for DPRQSum.

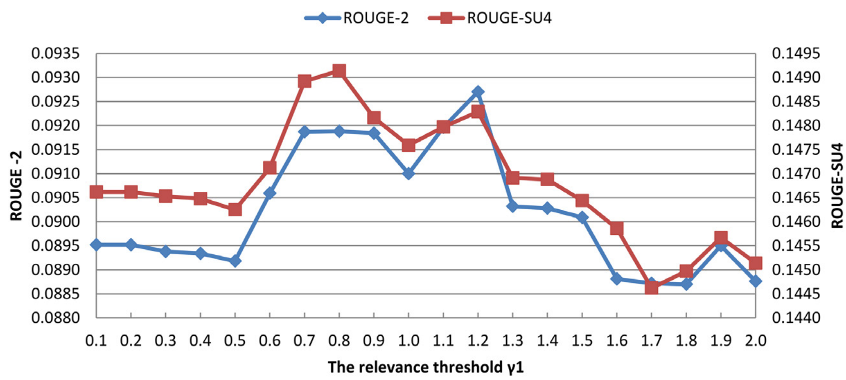



**Fig. 3.** ROUGE-2 and ROUGE-SU4 on DUC 2006 with $\gamma_1 = [0.1 - 2] * p_{avg}$ and $\gamma_2$ set to $\gamma_2 = 2 * n_{avg}$.

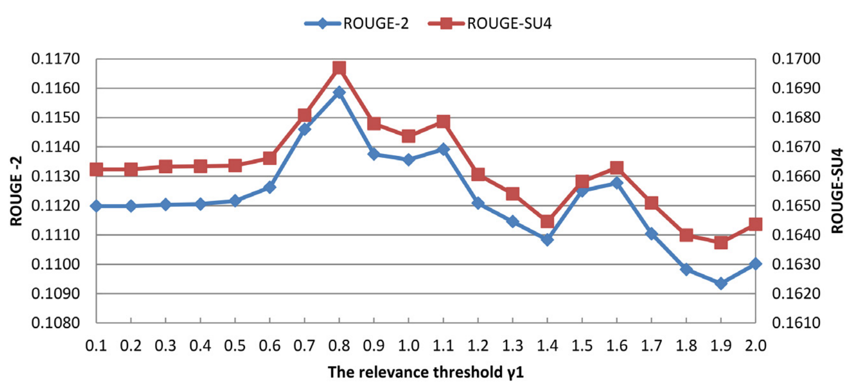



**Fig. 4.** ROUGE-2 and ROUGE-SU4 on DUC 2007 with $\gamma_1 = [0.1 - 2] * p_{avg}$ and $\gamma_2$ set to $\gamma_2 = 2 * n_{avg}$.

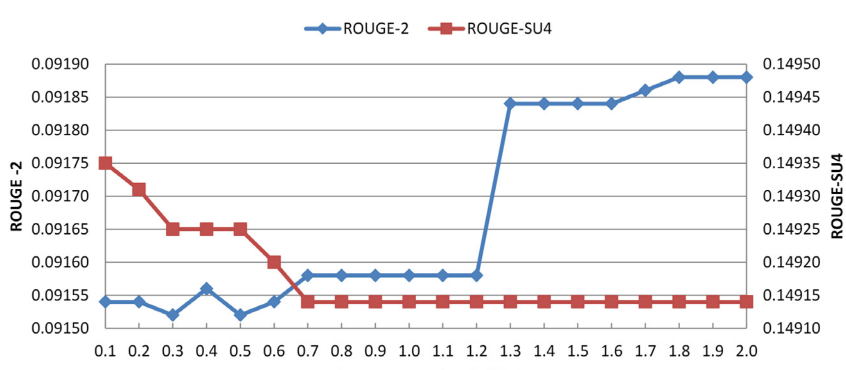



**Fig. 5.** ROUGE-2 and ROUGE-SU4 on DUC 2006 with $\gamma_2 = [0.1 - 2] * n_{avg}$ and $\gamma_1$ set to $\gamma_1 = 0.8 * p_{avg}$.

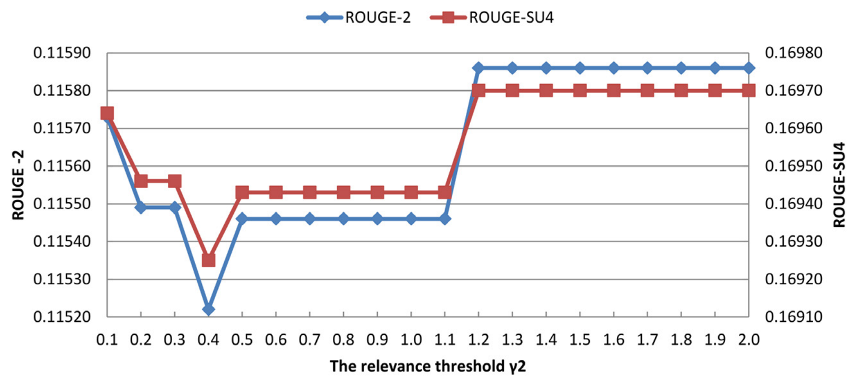



**Fig. 6.** ROUGE-2 and ROUGE-SU4 on DUC 2007 with $\gamma_2 = [0.1 - 2] * n_{avg}$ and $\gamma_1$ set to $\gamma_1 = 0.8 * p_{avg}$.

meets the requirement of topics coverage for query-focused MDS tasks, our method extracts a set of query relevant patterns together with background patterns from given documents to represent the query relevance of every sentence, so that not only the most informative sentences but also the most query relevant sentences can be selected to form a limited length summary to satisfy users' information needs. An innovative re-weighting process is applied on the derived query-relevant-patterns to punish patterns that are not biased towards the given query and maximise the representative power of query relevant patterns. By utilising well-established pattern mining, topic modelling and relevance ranking techniques together, the proposed method can meet the three major challenges of query-focused MDS: topic coverage, query relevance and balanced topical significance and query relevance. The experiments conducted on the DUC 2006 and DUC 2007 datasets verified the effectiveness and superior performance of the proposed query-focused MDS approach, DPRQSum.

## Acknowledgement

This paper was partially supported by Grant DP140103157 from the Australian Research Council, Australia.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2018.09.035.

## References

[1] R.M. Alguliev, R.M. Aliguliyev, M.S. Hajirahimova, Gendocsum + mclr: generic document summarization based on maximum coverage and less redundancy, Expert Syst. Appl. 39 (16) (2012) 12460–12473, http://dx.doi.org/10.1016/j.eswa.2012.04.067.

[2] R. Ferreira, L. de Souza Cabral, F. Freitas, R.D. Lins, G. de Frana Silva, S.J. Simske, L. Favaro, A multi-document summarization system based on statistics and linguistic treatment, Expert Syst. Appl. 41 (13) (2014) 5780–5787, http://dx.doi.org/10.1016/j.eswa.2014.03.023.

[3] E. Canhasi, I. Kononenko, Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization, Expert Syst. Appl. 41 (2) (2014) 535–543, http://dx.doi.org/10.1016/j.eswa.2013.07.079.

[4] J. Otterbacher, G. Erkan, D.R. Radev, Biased lexrank: passage retrieval using random walks with question-based priors, Inf. Process. Manage. 45 (1) (2009) 42–54.

[5] C. Cigir, M. Kutlu, I. Cicekli, Generic text summarization for turkish, in: Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on, 2009, pp. 224–229, http://dx.doi.org/10.1109/ISCIS.2009.5291848.

[6] J. Lin, N. Madnani, B.J. Dorr, Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 305–308.

[7] W. Luo, F. Zhuang, Q. He, Z. Shi, Exploiting relevance, coverage, and novelty for query-focused multi-document summarization, Knowl.-Based Syst. 46 (2013) 33–42, http://dx.doi.org/10.1016/j.knosys.2013.02.015.

[8] J. Tang, L. Yao, D. Chen, Multi-topic based query-oriented summarization., in: Proceedings of the 2009 SIAM International Conference on Data Mining, vol. 9, SIAM, 2009, pp. 1147–1158.

[9] A. Nenkova, K. McKeown, A survey of text summarization techniques, in: Mining Text Data, Springer, 2012, pp. 43–76.

[10] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, SIGMOD Rec. 22 (2) (1993) 207–216, http://dx.doi.org/10.1145/170036.170072.

[11] Y. Li, A. Algarni, M. Albathan, Y. Shen, M.A. Bijaksana, Relevance feature discovery for text mining, IEEE Trans. Knowl. Data Eng. 27 (6) (2015) 1656–1669, http://dx.doi.org/10.1109/TKDE.2014.2373357.

[12] Y. Li, A. Algarni, N. Zhong, Mining positive and negative patterns for relevance feature discovery, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 753–762.

[13] S.T. Wu, Y. Li, Y. Xu, Deploying approaches for pattern refinement in text mining, in: Proceedings of the Sixth IEEE International Conference on Data Mining, ICDM 2006, (ISSN: 1550-4786) 2006, pp. 1157–1161, http://dx.doi.org/10.1109/ICDM.2006.50.

[14] Y. Gao, Y. Xu, Y. Li, Pattern-based topics for document modelling in information filtering, IEEE Trans. Knowl. Data Eng. 27 (6) (2015) 1629–1642, http://dx.doi.org/10.1109/TKDE.2014.2384497.

[15] Y. Wu, Y. Li, Y. Xu, W. Huang, Mining topically coherent patterns for unsupervised extractive multi-document summarization, in: Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on, IEEE, 2016, pp. 129–136.

[16] Y. Wu, Y. Gao, Y. Li, Y. Xu, M. Chen, Mining topical relevant patterns for multi-document summarization, in: 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT, vol. 3, 2015, pp. 114–117, http://dx.doi.org/10.1109/WI-IAT.2015.136.

[17] H.P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (2) (1958) 159–165, http://dx.doi.org/10.1147/rd.22.0159.

[18] J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C.J. Pérez, Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach, Knowl.-Based Syst. 159 (2018) 1–8, http://dx.doi.org/10.1016/j.knosys.2017.11.029.

[19] X. Zhang, M. Lapata, F. Wei, M. Zhou, Neural latent extractive document summarization, 2018, arXiv preprint arXiv:1808.07187.

[20] L. Wang, H. Raghavan, V. Castelli, R. Florian, C. Cardie, A sentence compression based framework to query-focused multi-document summarization, 2016, arXiv preprint arXiv:1606.07548.

[21] R. Barzilay, K.R. McKeown, M. Elhadad, Information fusion in the context of multi-document summarization, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, 1999, pp. 550–557.

[22] P. Li, Z. Wang, W. Lam, Z. Ren, L. Bing, Salience estimation via variational auto-encoders for multi-document summarization., in: AAAI, 2017, pp. 3497–3503.

[23] K.R. McKeown, J.L. Klavans, V. Hatzivassiloglou, R. Barzilay, E. Eskin, Towards multidocument summarization by reformulation: progress and prospects, in: Proceedings of AAAI-99, 1999.

[24] J.M. Conroy, J.D. Schlesinger, D.P. O'Leary, Topic-focused multi-document summarization using an approximate oracle score, in: Proceedings of the COLING/ACL on Main Conference Poster Sessions, COLING-ACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 152–159.

[25] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Amer. Soc. Inform. Sci. 41 (6) (1990) 391.

[26] P. Fung, G. Ngai, One story, one flow: hidden markov story models for multilingual multidocument summarization, ACM Trans. Speech Lang. Process. 3 (2) (2006) 1–16.

[27] M.G. Ozsoy, I. Cicekli, F.N. Alpaslan, Text summarization of turkish texts using latent semantic analysis, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 869–876.

[28] L. Vanderwende, H. Suzuki, C. Brockett, A. Nenkova, Beyond sumbasic: task-focused summarization with sentence simplification and lexical expansion, Inf. Process. Manage. 43 (6) (2007) 1606–1618, http://dx.doi.org/10.1016/j.ipm.2007.01.023, Text Summarization.

[29] S. Ye, T.S. Chua, M.Y. Kan, L. Qiu, Document concept lattice for text understanding and summarization, Inf. Process. Manage. 43 (6) (2007) 1643–1662.

[30] A. Celikyilmaz, D. Hakkani-Tur, A hybrid hierarchical model for multi-document summarization, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 815–824.

[31] H. Daumé III, D. Marcu, Bayesian query-focused summarization, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 305–312, http://dx.doi.org/10.3115/1220175.1220214.

[32] A. Haghighi, L. Vanderwende, Exploring content models for multi-document summarization, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 362–370.

[33] D. Wang, S. Zhu, T. Li, Y. Gong, Multi-document summarization using sentence-based topic models, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 297–300.

[34] Y. Guangbing, W. Dunwei, Kinshuk, C. Nian-Shing, S. Erkki, A novel contextual topic model for multi-document summarization, Expert Syst. Appl. 42 (3) (2015) 1340–1352, http://dx.doi.org/10.1016/j.eswa.2014.09.015.

[35] Y. Djenouri, A. Belhadi, P. Fournier-Viger, Extracting useful knowledge from event logs: a frequent itemset mining approach, Knowl.-Based Syst. 139 (2018) 132–148, http://dx.doi.org/10.1016/j.knosys.2017.10.016.

[36] U. Yun, D. Kim, E. Yoon, H. Fujita, Damped window based high average utility pattern mining over data streams, Knowl.-Based Syst. 144 (2018) 188–205, http://dx.doi.org/10.1016/j.knosys.2017.12.029.

[37] N. Zhong, Y. Li, S.T. Wu, Effective pattern discovery for text mining, IEEE Trans. Knowl. Data Eng. 24 (1) (2012) 30–44, http://dx.doi.org/10.1109/TKDE.2010.211.

[38] Y. Xu, Y. Li, Generating concise association rules, in: Proceedings of the Sixteenth ACM conference on Conference on Information and Knowledge Management, ACM, 2007, pp. 781–790.

[39] X. Yan, H. Cheng, J. Han, D. Xin, Summarizing itemset patterns: a profile-based approach, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, pp. 314–323.

[40] S.T. Wu, Y. Li, Y. Xu, B. Pham, P. Chen, Automatic pattern-taxonomy extraction for web mining, in: Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on, IEEE, 2004, pp. 242–248.

[41] Y. Gao, Y. Xu, Y. Li, B. Liu, A two-stage approach for generating topic models, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 221–232.

[42] F. Xie, X. Wu, X. Zhu, Efficient sequential pattern mining with wildcards for keyphrase extraction, Knowl.-Based Syst. 115 (2017) 27–39, http://dx.doi.org/10.1016/j.knosys.2016.10.011.

[43] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, Multi-document summarization exploiting frequent itemsets, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, ACM, New York, NY, USA, 2012, pp. 782–786, http://dx.doi.org/10.1145/2245276.2245427.

[44] J.P. Qiang, P. Chen, W. Ding, F. Xie, X. Wu, Multi-document summarization using closed patterns, Knowl.-Based Syst. 99 (2016) 28–38, http://dx.doi.org/10.1016/j.knosys.2016.01.030.

[45] A.K. McCallum, MALLET: a machine learning for language toolkit, 2002, http://mallet.cs.umass.edu.

[46] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[47] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, 1994, pp. 487–499.

[48] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Discovering frequent closed itemsets for association rules, in: International Conference on Database Theory, Springer, 1999, pp. 398–416.

[49] M.J. Zaki, C.J. Hsiao, CHARM: an efficient algorithm for closed itemset mining, in: Proceedings of the 2002 SIAM International Conference on Data Mining, SIAM, 2002, pp. 457–473, http://dx.doi.org/10.1137/1.9781611972726.27.

[50] Y. Xu, Y. Li, G. Shaw, Reliable representations for association rules, Data Knowl. Eng. 70 (6) (2011) 555–575, http://dx.doi.org/10.1016/j.datak.2011.02.003.

[51] F. Wei, W. Li, Q. Lu, Y. He, Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, ACM, New York, NY, USA, 2008, pp. 283–290, http://dx.doi.org/10.1145/1390334.1390384.

[52] Y. Li, L. Zhang, Y. Xu, Y. Yao, R.Y.K. Lau, Y. Wu, Enhancing binary classification by modeling uncertain boundary in three-way decisions, IEEE Trans. Knowl. Data Eng. 29 (7) (2017) 1438–1451, http://dx.doi.org/10.1109/TKDE.2017.2681671.

[53] W. Pingxin, Y. Yiyu, CE3: a three-way clustering method based on mathematical morphology, Knowl.-Based Syst. 155 (2018) 54–65, http://dx.doi.org/10.1016/j.knosys.2018.04.029.

[54] Y. Yao, Three-Way decision: an interpretation of rules in rough set theory., in: International Conference on Rough Sets and Knowledge Technology, vol. 9, Springer, 2009, pp. 642–649.

[55] Y. Li, C. Zhang, J. Swan, An information filtering model on the web and its application in jobagent, Knowl.-Based Syst. 13 (5) (2000) 285–296, http://dx.doi.org/10.1016/S0950-7051(00)00088-5.

[56] S.E. Robertson, K.S. Jones, Relevance weighting of search terms, J. Assoc. Inform. Sci. Technol. 27 (3) (1976) 129–146.

[57] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 1998, pp. 335–336, http://dx.doi.org/10.1145/290941.291025.

[58] X. Wan, Using only cross-document relationships for both generic and topic-focused multi-document summarizations, Inf. Retr. 11 (1) (2008) 25–49, http://dx.doi.org/10.1007/s10791-007-9037-5.

[59] S. Yan, X. Wan, SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12) (2014) 2048–2058, http://dx.doi.org/10.1109/TASLP.2014.2360461.

[60] C.Y. Lin, Rouge: a package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, vol. 8, 2004.

[61] C.Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Vol. 1, Association for Computational Linguistics, 2003, pp. 71–78.

[62] M. Wasson, Using leading text for news summaries: evaluation results and implications for commercial summarization applications, in: Proceedings of the 17th International Conference on Computational Linguistics, vol. 2, Association for Computational Linguistics, 1998, pp. 1364–1368.

[63] J.M. Conroy, J.D. Schlesinger, J. Goldstein, D.P. O'leary, Left-brain/right-brain multi-document summarization, in: Proceedings of the Document Understanding Conference, DUC 2004, 2004.

[64] F. Wei, W. Li, Q. Lu, Y. He, A document-sensitive graph model for multi-document summarization, Knowl. Inform. Syst. 22 (2) (2010) 245–259, http://dx.doi.org/10.1007/s10115-009-0194-2.

[65] J.g. Yao, X. Wan, J. Xiao, Compressive document summarization via sparse optimization, IJCAI (2015) 1376–1382.

[66] R. He, J. Tang, P. Gong, Q. Hu, B. Wang, Multi-document summarization via group sparse learning, Inform. Sci. 349 (2016) 12–24, http://dx.doi.org/10.1016/j.ins.2016.02.032.

[67] J. Jagarlamudi, P. Pingali, V. Varma, Query independent sentence scoring approach to duc 2006, in: Proceeding of Document Understanding Conference, DUC-2006, 2006.

[68] R.K. Prasad Pingali, V. Varma, Iiit hyderabad at duc 2007, in: Proceedings of DUC 2007, 2007.

[69] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.