Learning from summaries: supporting e-learning activities by means of document summarization

Learning From Summaries: Supporting e-Learning Activities by means of Document Summarization

Elena Baralis and Luca Cagliero

Dipartimento di Automatica e Informatica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: *name.surname@polito.it*

Keywords E-learning, Document Summarization.

Abstract *E-learning platforms allow users with different skills to explore large collections of electronic documents and annotate them with notes and highlights. Generating summaries of these document collections is potentially useful for gaining insights into teaching materials. However, most existing summarizers are general-purpose. Thus, they do not consider neither annotations nor user skills during the document summarization process.*

This paper studies the application of a state-of-the-art summarization system, namely the Itemset-based Summarizer (ItemSum), in an e-learning context. The summarizer produces an ordered sequence of key phrases extracted from a teaching document. The aim of this work is threefold: (i) Evaluate the usefulness of the generated summaries for supporting individual and collective learning activities in a real context, (ii) understand to what extent document highlights, annotations, and user skill levels can be used to drive the summarization process, and (iii) generate multiple summaries of the same document tailored to users with different skill levels. To accomplish Task (i), a crowd-sourcing experience of evaluation of the generated summaries was conducted by involving the students of a B.S. course given by a technical university. The results show that the automatically generated summaries reflect, to a large extent, the students' expectations. Hence, they can be useful for supporting learning activities in university-level Computer Science courses. To address Task (ii), three extended versions of the ItemSum summarizer, driven by highlights, annotations, and user skill levels, respectively, have been proposed and their performance improvements with respect to the baseline version have been validated on benchmark documents. Finally, to accomplish Task (iii) multiple summaries of the same benchmark documents have been generated by considering only the annotations made by the users with a different skill level. The results confirm that the summary content reflects the level of expertise of the targeted users.

1 Introduction

Recent advances in ICT technologies offer great opportunities to improve the quality of learning activities. E-learning platforms allow learners to access teaching materials through different devices, such as laptops, smartphones, and tablets Moore et al. [2011]. In the meanwhile, the evolution of Internet-based applications has opened great opportunities to ease remote teacher-learner interactions. Hence, on the one hand, teachers can exploit web-based learning tools to share teaching materials with their learners and to publish assignments, exam grades, and questionnaires Hossain et al. [2014]. On the other hand, learners can remotely explore electronic books, notes, and slides and annotate them with notes or highlights Zoubib and Jali [2014], Ibanez et al. [2014].

In most learning contexts, learners have to explore a large number of potentially long electronic textual documents. For example, to prepare for academic exams, students have to read electronic books, lecture notes, and scientific articles. To gain insights into these potentially large document collections, e-learning systems should integrate advanced data analysis and mining tools Tan et al. [2005].

Sentence-based document summarizers are automated tools aimed to extract salient information from textual electronic documents. State-of-the-art summarizers rely on data mining or information retrieval techniques, such as classification Li et al. [2009], clustering Wang et al. [2011], graph mining Thakkar et al. [2010], and pattern Baralis et al. [2012] and lexical chain Pourvali and Abadeh [2012] mining. In e-learning systems, summaries automatically generated by teachers/learners in the past can be shared, accessed by other learners/teachers with different level of expertise, and possibly updated as soon as new teaching content becomes available. For example, university-level materials from past academic years can be explored, reused, and updated by the currently enrolled students, by the professors, or by the teaching assistants. Similarly, students may summarize the notes taken by the other students of the same course and shared on the e-learning platform to validate the content of their personal notes. In general, learners may access the automatically generated summaries (i) prior to document exploration, to have a preliminary idea of the document content, (ii) after reading documents, to revise the lesson learnt, or (iii) in place of the original documents, to overcome time constraints or bandwidth limitations (e.g., in mobile learning Yang et al. [2012]). Unfortunately, many teaching documents have no abstract. Exploiting automatic document summaries to support teaching activities is an appealing research direction, which, to the best of our knowledge, has never been addressed in literature.

Teaching documents accessible through e-learning systems are often enriched with additional information. For example, users with different skill levels may annotate documents with notes or highlights. Highlights (e.g., underlined text, circled text), annotations (e.g., notes written alongside of the text), and skill levels of the users annotating the text (e.g., beginner, expert) represent additional information on the teaching documents that is worth considering to summarize teaching documents. However, general-purpose summarizers are unable to consider such information during document summarization. A more detailed overview of the state-of-the-art is given in Section 2.

As an application example, let us consider a university-level professor, who exploits electronic documents in support of her oral lectures. During lectures the teacher annotates the documents with textual notes and highlights. In the meanwhile, students access the electronic version of the documents through their tablets or laptops and enrich them with highlight and notes based on their common knowledge and on their comprehension of the teacher's explanations. Hence, documents acquired from e-learning platforms are enriched with notes/highlights made by users with different skill levels. Shared materials annotated during past academic years or shared by learners/students of the current academic year can be summarized to support individual and collective learning activities. The exploitation of highlights, annotations, and user skills in teaching document summarization opens the following research issues: (i) Does document summarization driven by additional information produce higher-quality summaries than traditional approaches exclusively based on the original document content? (ii) Can document summarizers automatically generate multiple summaries of the same document tailored to users with different skill levels?

This paper investigates the applicability of a state-of-the-art text summarization system, called Itemset-based Summarizer (ItemSum) Baralis et al. [2012], in an e-learning context. The baseline summarizer version takes as input a teaching document, possibly partitioned into sections, subsections, and paragraphs, and it produces a succinct yet informative summary consisting of the most representative document sentences. Document summaries are provided to learners as additional material in support of study and revision. The applicability of the proposed approach depends on the type of analyzed documents. It appears to be particularly useful when (i) dealing with teaching materials (e.g., lecture notes, book chapters, technical reports) that do not contain any abstract or outline, (ii) readers need to quickly access the key information hidden in the original document before reading them or during revision, or (iii) the devices used to explore the teaching content have bandwidth limitations (e.g., in mobile learning Yang et al. [2012]). To demonstrate the applicability of document summarization in a real e-learning context, a crowd-sourcing experience of evaluation of the summary generated from a real teaching document was conducted. The activity involved the students of a B.S. Database course given by a technical university. A popular scientific article Page et al. was given as additional teaching material. Students were asked to fill a questionnaire to select the top-5 key phrases from the given article. Based

on the students' responses, a golden summary was generated and compared with the automatically generated summary. To generate the article digest using the ItemSum algorithm, the article was divided into sections. Thus, the relative importance of each document sentence was evaluated not only globally on the whole article but also locally within each section. The achieved results confirm that the automatically generated summaries reflect, to a large extent, the student's expectations, because three out of five sentences are in common between the automatically generated and the golden summaries. Hence, the generated summaries appear to be useful for supporting learning activities in university-level Computer Science courses.

To effectively cope with teaching documents enriched with additional information, this paper also proposes three new extended versions of the ItemSum summarizer that respectively consider highlights, annotations, and user skill levels to drive the document summarization process. The newly proposed summarizers rely on different sentence ranking functions based on the type of additional information they consider beyond the original document content. Specifically, the following ranking functions are considered: (i) *Ranking based on highlights*, which considers the presence of highlighted sentences and annotations to accurately select the most representative sentences to include in the summary, while disregarding the annotation content and the user skill levels. (ii) *Ranking based on the annotation content*, which considers not only the presence of an annotation, but also its content, while disregarding the user skill levels. (iii) *Ranking based on user skill levels*, which considers both the annotation content and the user skill levels.

The experiments, performed on benchmark documents Document Understanding Conference [2004], show that driving document summarization with highlights, annotations, and user skill levels improves the quality of the generated summaries compared to traditional approaches exclusively based on the original document content. Finally, on the same documents we generated multiple summaries by considering, for each run, only the annotations made by the users with a different skill level. The achieved results confirm that the summarizer can automatically generate summaries tailored to different categories of users, because the most technical content occurs only in the summaries provided to highly skilled users.

This paper is organized as follows. Section 2 compares the proposed work with state-of-the-art related approaches. Section 3 presents and thoroughly describes the context under analysis and the newly proposed summarization strategy, while Section 4 experimentally evaluates the effectiveness of the proposed approach as well as its applicability in a real-life context. Finally, Section 5 draws conclusions and presents future developments of this work.

2 Related work

Extractive document summarization entails selecting most significant document content from a (set of) textual document(s) Tan et al. [2005]. Extractive summarization algorithms can be categorized as follows: *sentence-based*, if they divide the document content into sentences and pick the most informative sentences Conroy et al. [2011], Baralis et al. [2013a], or *keyword-based*, if they extract salient keywords summarizing the document content Dredze et al. [2008], Lin and Hovy [2003]. A complementary classification is the following: *multi-document*, if they produce one single summary of multiple documents, or *single-document*, if they generate one summary per document. This paper addresses the problem of integrating a sentence-based multi-document summarizer in an e-learning environment. Since different sections of the same document may cover different facets of the same topic, each section is first treated as a distinct input item, then one summary per document is generated.

To effectively perform sentence selection, sentence-based summarizers adopt different techniques: (i) Clustering, (ii) graph ranking, (iii) optimization strategies, and (iv) frequent itemset mining. *Clustering*-based approaches (e.g., Wang and Li [2010], Wang et al. [2011]) exploit clustering algorithms to group similar sentences and then pick most significant sentences within each group. *Graph*-based approaches (e.g., Thakkar et al. [2010], Baralis et al. [2013b]) construct a graph whose nodes represent document terms, while edges connect pairs of nodes and they are weighted by pairwise node similarity measures. *Optimization* strategies perform Singular Value Decomposition (SVD) Steinberger et al. [2011], Integer Linear Programming Gillick et al. [2009], or similar strategies to extract salient document sentences. In Conroy et al. [2011] optimization strategies are combined with Hidden Markov Models to address the multilingual document summarization problem. *Itemset*-based approaches (e.g., Hynek and Jezek [2003], Baralis et al. [2012]) exploit frequent itemsets, which represent sets of document terms of arbitrary length, to capture the underlying correlations among multiple terms. Recently, an highly effective summarizer based on frequent itemsets, namely Itemset-based Summarizer (ItemSum), has been proposed Baralis et al. [2012]. To consider only the most informative and non-redundant patterns hidden in the analyzed data, itemsets are first extracted using the entropy-based strategy Mampaey et al. [2011]. Then, a greedy strategy is adopted to select the subset of sentences that best covers the extracted itemsets. ItemSum was tested on benchmark Document Understanding Conference [2004] and real news document collections. On these documents it appears to perform significantly better than state-of-the-art approaches. This paper inves-

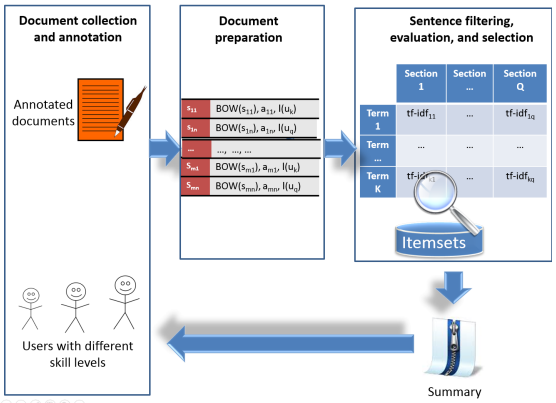

Figure 1: The Learning From Summaries architecture.

tigates the application of the ItemSum algorithm in an e-learning context. Instead of summarizing multiple documents in one single summary, it partitions documents into sections, considers each section as a separate input, and then generates *one summary per document*. Furthermore, since documents acquired from e-learning platforms are commonly enriched with annotations made by users with different skill levels, it proposes three extended versions of the ItemSum algorithm that consider also highlights, annotations, and user skill levels during the summarization process. In addition, the problem of generating multiple summaries of the same document tailored to users with different skill levels has also been addressed.

A parallel issue is to consider the document structure during summarization. Some previous works (e.g., Binh Tran [2013], Yan et al. [2011]) focus on extracting important events or sentences and putting them in a chronological time span. For example, in Binh Tran [2013] the authors exploit the chronological characteristics of online news to highlight the most important events in temporal order thus improving summary readability. Other approaches (e.g., Gross et al. [2014]) consider the order of appearance of the sentences in the documents because, for example, in news articles key information is likely to appear in the first document sentences. Still others (e.g., Baralis et al. [2015b], Li et al. [2009]) tailor the summarization process to heterogeneous/homogeneous content, respectively, by exploiting ad hoc term evaluators (tf-idf vs. tf-df) and objective functions (coverage vs. diversity). Unlike Binh Tran [2013], Li et al. [2009], Yan et al. [2011], Gross et al. [2014], Baralis et al. [2015b], the approach proposed in this work is not focused on news summarization. Hence, the chronological order of events is, in general, not available and the order of appearance of the sentences in the documents may not matter. To make the proposed approach as much general as possible, we select the sentences that are deemed as most significant across different sections of the same document.

Some preliminary research efforts have been devoted to integrating summarization algorithms into e-learning tools. For example, since learners often perform the same queries on the online document collection, an interesting research problem is to generate summarized answers to most frequent queries to avoid the need for maintaining a huge knowledge base thus improving e-learning system efficiency Saraswathi et al. [2011]. A complementary research problem is to make learning material accessible through mobile devices. In Yang et al. [2012] a personalized text-based content summarizer is presented. It relies on abstractive text summarization techniques, which are exploited to automatically infer personalized summaries suitable for mobile learning. The aim of this work is radically different with respect to Saraswathi et al. [2011], Yang et al. [2012] because the newly proposed approach is not query-driven and is extractive and not abstractive, i.e., it selects existing document content as part of the resulting summary rather than inferring new content from the input text. A preliminary version of this work was presented in Baralis et al. [2015a]. The paper presented the results of a summary evaluation experience in an e-learning context. Summaries were generated using the ItemSum algorithm Baralis et al. [2012]. In Baralis et al. [2015a] the analyzed documents are not enriched with highlights, annotations, and user skill levels. Hence, the document summarization process is neither driven by annotations/highlights nor tailored to users with different levels of expertise. The summarizer extensions proposed in this paper are able to cope with learning documents enriched with additional information. Such additional information cannot be handled by the ItemSum algorithm. To the best of our knowledge, this paper is the first attempt to consider document highlights, annotations, and user skills during summary generation.

3 Learning From Summaries

Learning From Summaries (LFS) is a new data mining approach to exploiting document summarization techniques in support of learning activities. The main steps of the proposed approach, depicted in Figure 1, are briefly summarized below. A more detailed description of each step is given in the following sections.

Document collection and annotation. During learning activities users annotate documents with notes or highlights. Users accessing the documents are characterized by a role (e.g., teacher, student, reviewer), which, in turn, can be mapped to specific skill levels between zero and one (e.g., beginners level 0.1, experts level 1). The annotated documents are collected and stored into a centralized data repository (see Section 3.1).
Document preparation. Documents are prepared to the next summarization process. Specifically, each document is partitioned into sections to differentiate between different parts of the text during the summarization process. Only the textual parts of the documents are considered. To tailor data to the itemset mining process the preprocessed sentences are transformed into a transactional data format. Furthermore, stemming and stopword algorithms are applied to improve the quality of the summarization process (see Section 3.2).
Document summarization. Based on the type of additional information available (highlights, annotations, or user skill levels), different extended versions of the Itemset-based Summarizer (ItemSum) Baralis et al. [2012] are applied to each document of the collection. The summarizers take as input one document at a time, possibly partitioned into sections and enriched with additional information, and generate a summary per document. If the information about the user skill levels is available, the summarizers allow generating also multiple summaries per document, each one tailored to users with different levels of expertise (see Section 3.3).

Currently, we developed a tool prototype, for research purposes only, where document preparation and summarization are performed semi-automatically according to the user-specified input parameters. A more thorough description of each step is given below.

3.1 Document collection and annotation

E-learning platforms allow learners to share textual documents for teaching purposes through different devices, such as laptops, smartphones, and tablets Moore et al. [2011]. Documents are uploaded into a shared repository and accessed by users with different levels of expertise.

Users of e-learning systems commonly have a role, based on which specific actions are granted (e.g., evaluate assignments, create questionnaires, post comments, upload exam simulations). User roles can be straightforwardly mapped to skill levels by domain experts (e.g., beginner users have level 0.1, expert users have level 1). For example, let us consider a learning platform used to manage the activities related to an university-level course. In this context, professors, teaching assistants, and students are examples of users. According to their level of expertise, professors can be classified as *experts* thus a skill level equal to 1 is assigned, teaching assistants have skill level 0.5 (*fair*), whereas *beginner* students have skill level 0.1. Professors upload lecture notes and other textual documents on the educational Web portal, send messages to students through the internal communication system, create questionnaires and tests, and evaluate the student assignments. Teaching assistants create and modify questionnaires and evaluate the student assignments. Students explore learning documents, annotate them, post comments/questions on the portal, and upload their solutions of the given assignments.

Users can annotate documents with notes or highlights. Notes are either rephrases of part of the text or extra textual content giving more details on a specific subject. Highlights are graphical signs exploited to mark part of the text (e.g., marked text, underlined text, circled text). Annotations can be made by users with different skill levels. For example, to stress the importance of specific concepts during the oral lessons professors can highlight some portions of the text or they can annotate the document content with short phrases or keywords clarifying the underlying information. In the meanwhile, students can highlight parts of the text which they deem as particularly relevant based either on their common knowledge or on the level of comprehension of the professor's explanations. After the lesson, students can read again the documents and further annotate them with text rephrases or additional comments. As discussed in the following, highlights/annotations made by users in the past can be used to drive document summarization.

Hereafter we will denote as $\mathbf{D}$ the collection of textual documents available in the learning system. Each documents $D_i \in \mathbf{D}$ consists a set of sentences s_i^j, clustered into different sections. Documents are annotated by users with different skill levels. For the sake of simplicity, we will map annotations to document sentences, i.e., each sentence s_i^j is enriched with the corresponding annotation a_i^j, Hence, we assume that annotations referring to a part of a sentence can be approximately mapped to the entire sentence. For each annotation a_i^j the user u_k who annotated the sentence and the corresponding skill level $l(u_k)$ are known. For example, Table 1 contains an example of annotated document, consisting of two different sections (A and B). Sections are composed of three sentences each. Document sentences are annotated by users with different skill levels. For instance, sentence with Id 3 is annotated with comment *"Large" may potentially mean Gigabytes or Terabytes of data!* by user Alice, whose skill level is 0.5 (*fair*).

Note that sentences may have no annotation (e.g., sentence with Id 1) or may be enriched with multiple annotations (possibly made by users with different skill levels).

The document collection, possibly annotated with highlights, annotations, and user skill levels, is integrated into a unique repository to allow data preprocessing and summarization.

Table 1: Example of annotated document d_1.

Id	Sentence	Annotation	User	Skill
		Section A		
1	This is a short introduction to data mining.	-	-	-
2	Data mining is a subfield of computer science.	It is related to machine learning and artificial Intelligence	Bob	1 (Expert)
3	It analyzes large datasets.	"Large" may potentially mean Gigabytes or Terabytes of data!	Alice	0.5 (Fair)
		Section B		
4	Data mining algorithms are classified as supervised and unsupervised	-	-	-
5	Supervised algorithms focus on predicting the value of a variable	Classification, regression	Alice	0.5 (Fair)
6	Unsupervised ones aim at representing significant patterns hidden in the analyzed data	Patterns to be discovered	John	0.1 (Beginner)

3.2 Document preparation

This block processes the document collection **D**, possibly enriched with annotations, highlights, and user skills, and prepares the textual content of the raw documents to the next summarization process (see Section 3.3). To automate the preprocessing phase, most preprocessing steps are applied semi-automatically based on the user's specifications. Most relevant preprocessing steps are enumerated below.

Non-textual content filtering. Documents usually contain also non-textual content (e.g., images, videos etc.). Since the proposed approach specifically addresses textual document summarization, non-textual content is automatically filtered out before running the summarization process. Note that since parallel research efforts have already addressed the transformation of non-textual learning content into textual form (e.g., Chang et al. [2011]), other existing analytical tools might be integrated to cope with multimedia content.

Document splitting. Documents are commonly structured into sections and subsections. Each section covers a different facet of the main document subject. To take the document structure into account during the summarization process, the document structure is automatically detected and partitioned into sections and each section of the original document will be considered as a separate input document of the next summarization step. Splitting documents into sections allows the summarizer to evaluate the importance of each sentence both locally within the corresponding section and globally in the whole document. Furthermore, documents are split into sentences based on punctuation, i.e., two phrases separated by full stop, question mark, colon, or semicolon are considered as distinct sentences.

Stemming and stopword elimination. To effectively process textual documents, text mining algorithms often reduce document words to their base or root form (i.e., the stem) Tan et al. [2005]. Furthermore, frequently occurring but not informative words, i.e., the stopwords, are usually filtered out before text processing. Examples of stopwords are articles, prepositions, and conjunctions. The aforesaid preprocessing steps are particularly useful for summarization purposes, because since the sentence evaluation process relies on word frequency counts (see Section 3.3) considering also stopwords and non-stemmed words may bias the process of evaluation of the document sentences thus yielding low-quality summaries. For example, sentences containing many articles and prepositions may be wrongly included in the summary regardless of the importance of the remaining content. Currently, the system integrates the Wordnet stemming and stopwords algorithms for English-written documents (http://wordnet.princeton.edu). However, to cope with documents written in different languages, different stemming and stopword elimination algorithms can be straightforwardly integrated as well.

Data transformation. After applying stemming and stopword elimination, each document sentence can be modeled as a list of (possibly repeated) stems, called BOW representation Tan et al. [2005]. Specifically, the Bag-Of-Word (BOW) sentence representation consists of the collection of sentence stems generated by disregarding grammar and even word order but keeping multiplicity. Since most itemset mining algorithms rely on a transactional data format Tan et al. [2005], starting from the BOW representation a transformed version of each analyzed document is generated. It consists of a set of transactions rather than a set of BOWs. Each transaction corresponds to a different document sentence and consists of the set of not repeated stems contained in the preprocessed sentence.

3.3 Document summarization

The summarization algorithm takes as input the preprocessed document collection **D**, possibly enriched with annotations and user skills, and it generates one summary per document in **D**. Since the level of enrichment of the documents acquired from the e-learning system can be different, the newly proposed summarizer adopts different sentence ranking strategies according to the characteristics of the input data. More specifically, the

following rankings are considered: (i) *Baseline ranking*, which considers the presence of highlighted sentences and annotations to accurately select the most representative sentences to include in the summary. It disregards the annotation content and the user skill levels. This ranking function can be applied to documents with highlighted sentences. (ii) *Content-based ranking*, which considers not only the presence of an annotation, but also its content, while disregarding the user skill levels. This ranking function can be applied to documents with textual notes (e.g., notes, rephrases, keywords). (iii) *Skill-based ranking*, which considers both the annotation content and the user skill levels. This ranking function can be applied to documents annotated with highlights, textual annotations, and user skill levels.

To effectively summarize the enriched learning documents, learners should choose the most appropriate ranking function based on the additional information available on the input documents. The process of summarization of each document of the collection entails the following steps:
(i) Itemset mining. A model consisting of a subset of highly informative itemsets is extracted from the transactional version of the input document.
(ii) Sentence filtering. Document sentences not covered by any frequent itemset are filtered out because they are unlikely to represent salient information.
(iii) Sentence evaluation and selection. The remaining sentences of the document are ranked according to their content and annotations and included in the summary.
A more thorough description of each step is given below.
Itemset mining. In our context, a k-itemset (i.e., an itemset of length k) is a set of (not repeated) stems. The frequency of occurrence (support) of an itemset in a document is the number of sentences in which it occurs. Frequent itemset mining entails discovering all itemsets (of arbitrary length) whose support is above a given (analyst-provided) threshold *minsup*.

For example, recalling the example dataset in Table 1 both stems *Data* and *Mine*[1] occur in the preprocessed document version. Since these stems co-occur in half of the document sentences (i.e., in sentences with ids 1, 2, and 4) the support of itemset { *Data*, *Mine*} is $\frac{3}{6}$, i.e., 50%.

Since the set of all frequent itemsets mined from a transactional data is potentially redundant, many research efforts have been devoted to discovering top-K informative yet non-redundant itemsets (e.g., Jaroszewicz and Simovici [2004], Mampaey et al. [2011]). Similar to ItemSum Baralis et al. [2012], we adopt an advanced algorithm for non-redundant itemset mining Mampaey et al. [2011] which relies on an entropy-based heuristics.
Sentence filtering. Since frequent itemsets represent the most salient information hidden in the analyzed document, we exploit the mined itemsets to select the most interesting document sentences. Specifically, only the sentences that are covered by at least one extracted itemset are kept, because they are most likely to cover interesting knowledge. On the other hand, sentences not including any frequent combination of words are deemed as not appropriate to appear in the summary, because they may represent irrelevant or marginal information.

Sentence evaluation and selection. This step entails ranking the remaining sentences based on their content and annotations. Top ranked sentences are placed first in the summary, because they best represent the document content. Depending on the available information (e.g., highlights, notes, user skill levels), sentence ranking relies on different strategies. Hereafter we will introduce the sentence ranking strategies adopted by the proposed summarizer.
Baseline ranking. This ranking strategy considers documents whose sentences are enriched with highlights only, i.e., when sentences are labeled as annotated or not, but neither textual annotations nor user skills are available. Let $\text{Ann}(s_i^j)$ be a boolean function returning value 1 if sentence s_i^j is highlighted, 0 otherwise. Let $\text{Tf-idf}(s_i^j)$ be the average term frequency-inverse document frequency (tf-idf) Tan et al. [2005] value of the stems in sentence s_i^j. The score $\text{Rank}(s_i^j)$ assigned to sentence s_i^j is computed as follows:

$$\text{Rank}(s_i^j) = \text{Tf-idf}(s_i^j) \cdot e^{\text{Ann}(s_i^j)}$$

The sentence rank is given by the the product of two complementary terms: (i) a term evaluator and (ii) a highlight indicator. Sentence terms are evaluated by combining the tf-idf values of the corresponding stems. The tf-idf evaluator Lin and Hovy [2003] is an established and widely used statistics that is intended to reflect how important a term is in a document of a collection or corpus. The key idea behind the tf-idf statistics is that terms appearing frequently in a few sections of the document (i.e., high local term frequency), but rarely in the whole document (i.e., low document frequency), are the most effective ones in discriminating among sentences in a collection. A more detailed description of the tf-idf statistics is reported in Tan et al. [2005].

Expression $e^{\text{Ann}(s_i^j)}$ is a boolean penalty score used to discriminate between annotated sentences and not. Specifically, annotated sentences get maximal score (e), whereas non-annotated sentence get minimal score (1).
Content-based ranking. This ranking strategy handles documents whose sentences are enriched with textual annotations, but it does not consider the level of expertise of the users annotating the sentences. Let $\text{Tf-idf}(s_i^j)$

[1] This stem was generated from *Mining*.

and Tf-idf(a_i^j) be the average tf-idf values of the stems in sentence s_i^j and annotation a_i^j, respectively. If sentence s_i^j has multiple annotations Tf-idf(a_i^j) is averaged over all the annotations corresponding to s_i^j.

The score Rank(s_i^j) assigned to sentence s_i^j is given by:

$$\text{Rank}(s_i^j) = \alpha \cdot \text{Tf-idf}(a_i^j) \cdot e^{\text{Ann}(s_i^j)} + (1 - \alpha) \cdot \text{Tf-idf}(s_i^j)$$

where α is an analyst-provided parameter ranging between zero and one. The ranking function is given by the sum of two complementary terms: (i) a sentence evaluator and (ii) an annotation evaluator. Both evaluators rely on the tf-idf statistics and are computed on the corresponding stems. The weighting factor α is used to privilege the importance of user annotations with respect to the document content or vice versa. The higher α we set the more discriminative user annotations are during the sentence ranking process. If α is set to 0 annotations are disregarded during the sentence ranking process. Conversely, if α is set to 1 the content of the document is ignored. An empirical evaluation of the impact of parameter α on the summarizer performance is given in Section 4.

Skill-based ranking. This ranking strategy handles documents whose sentences are enriched with both textual annotations and user skill levels. Let a_i^j be the annotation made by user u_k to sentence s_i^j and let $l(u_k)$ be the skill level of user u_k. If sentence s_i^j has multiple annotations, $l(u_k)$ is the average skill level of the users annotating s_i^j. Without any loss of generality, hereafter we will assume $0 \leq l(u_k) \leq 1$.

The score Rank(s_i^j) assigned to sentence s_i^j is given by:

$$\text{Rank}(s_i^j) = \alpha \cdot \text{Tf-idf}(a_i^j) \cdot e^{\text{Ann}(s_i^j) \cdot l(u_k)} + (1 - \alpha) \cdot \text{Tf-idf}(s_i^j)$$

The formula above is a generalization of the former ones. Unlike the *Content-based* ranking, this ranking function gives different importance to the annotations based on the user skill levels. Specifically, the tf-idf score of the annotations is weighted by the corresponding user skill level. Hence, the annotations made by lowly skilled users are, on average, characterized by a lower rank.

The summary of each document consists of the subset of selected sentences ranked according to the chosen ranking strategy.

3.4 Tailoring summaries to different user skills

Applying the summarization process on the whole document collection generates a unique summary for all end users. This choice could be suboptimal, because users with different skill level may expect to read different summaries according to their level of expertise. For example, summaries provided to beginner users should be less technical than those provided to expert learners.

If documents are enriched with user skills, the summarization tool allows, as an option, to tailor summaries to users with different level of expertise. To tailor summaries to different user skill levels, the summarization process is run multiple times on the same document. For each run, only the annotations made by the users with a given skill level are considered. Multiple runs related to different skill levels produce multiple summaries of the same document tailored to users different levels of expertise.

4 Experiments

We conducted experiments on real and benchmark documents to answer to the following questions:
(i) Do the summaries generated by the baseline ItemSum summarizer version Baralis et al. [2012] from real teaching documents meet the expectations of real students of a university-level Computer Science course?
(ii) Are the automatically generated summaries comparable with those given by the document authors (e.g., the document abstract) or manually generated by domain experts?
(iii) Is the summarization process driven by document highlights, annotations, and/or user skill levels more effective than those exclusively based on the document content?
(iv) Do the summaries generated from the same benchmark document and tailored to different user skills meet the users' expectations?

To address issues (i) and (ii), we validated the usefulness of the proposed approach in a real application scenario, i.e., the summarization of teaching materials of an university-level Computer Science course, by conducting a crowd-sourcing experience of evaluation of the summaries generated by the baseline document summarizer. The evaluation allowed us to qualitatively evaluate the level of satisfaction of real system users (i.e., the students of a B.S. Computer Science course) on the generated summaries (see Section 4.1).

To address issues (iii) and (iv), we tested our approach on benchmark news documents, i.e., the Document Understanding Conference 2005 (DUC'05) SCU-marked collections. Specifically, to address issue (iii) we quantitatively compared the performance of different sentence ranking strategies according to standard evaluation scores Lin and Hovy [2003] (see Section 4.2). To address issue (iv) we partitioned the annotations based on the

skill level of the corresponding user and then we qualitatively compared the summaries generated by driving the summarization process with different annotation clusters.

All the experiments were performed on a quad-core 3.30 GHz Intel Xeon workstation with 16 GB of RAM, running Ubuntu Linux 12.04 LTS. For the baseline version of ItemSum algorithm Baralis et al. [2012] and the entropy-based itemset extractor Mampaey et al. [2011] we exploited the implementations provided by the respective authors. Unless otherwise specified, hereafter we will consider the standard algorithm configuration reported in Baralis et al. [2012].

4.1 Qualitative summary evaluation on real documents: a crowd-sourcing experience

This section aims at demonstrating the applicability and usability of sentence-based summarization in a real e-learning context, i.e., research issues (i) and (ii). To address these issues, we performed a crowd-sourcing experience of evaluation of the summaries generated from a real teaching document given to the students of an university-level Computer Science course. The experience, carried out on a voluntary basis, lasted three weeks and was conducted by involving 48 students of a Database course (a 2nd year B.S. course) given by the Politecnico di Torino, an Italian technical university. Students were invited to participate to the crowd-sourcing experience by filling an anonymous questionnaire available at the course website. The activity consisted in reading a scientific article Page et al., which presents the key ideas behind the popular PageRank indexing algorithm, and selecting at most 5 key phrases from the given article based on their personal judgment and experience. The article was given to the students as additional teaching material and was deemed as an example of technical document whose summary could be useful for supporting study and revision. Based on student preferences, a golden summary was generated by ranking the article sentences in order of decreasing preference (i.e., from the top preferred sentence to the least preferred one) and selecting the top ranked sentences.

To generate the automatic summary, images, tables, graphs, references, and sentences including formulas or referring to mathematical proofs were not considered because the summarizer exclusively copes with textual content. To perform a fair evaluation, we first removed the abstract from the copy of the article given to the students, because it already represents an example of document summary. In Section 4.1.2, the abstract of the article will be compared with both the automatically generated and the golden summaries. To select the most important article sentences to include the summary by considering also the importance of each sentence locally within each section, the article content was preliminary partitioned into sections according to the following schema chosen by the article authors: (i) Introduction and Motivation. (ii) A ranking for every page of the Web. (iii) Implementation. (iv) Convergence properties. (v) Searching with PageRank. (vi) Personalized PageRank. (vii) Applications. (viii) Conclusion.

4.1.1 Comparison between golden and automatically generated summaries

Table 2 summarizes the results of the crowd-sourcing experience of evaluation. It reports the top-5 sentences selected by the summarizer as well as all the other sentences that received more than five preferences by the students. Column *Sum. rank* indicates the rank of the sentence in the automatically generated summary. Columns *Prefs.* and *Pref. rank* indicate, for each sentence, the number of preferences given by the students and the corresponding rank according to the distribution of the number of preferences, respectively. In Table 2 the sentences of the summary are sorted in order of decreasing relevance based on the baseline ranking criterion (see Section 3.3), while the other sentences are sorted in order of decreasing student preference.

The top-2 sentences selected by the summarizer appear to be the top preferred ones according to the crowd-sourcing evaluation (19 preferences each). The third and fourth sentence received 13 and 9 preferences thus they placed third and seventh, respectively. Notably, the sentence ranking in the summary corresponds to the preference rank achieved in the crowd-sourcing experience. Hence, the sentence selection process reflects the students' expectations. In summary, three out of five sentences selected by the document summarizer matched the user expectations, because they are in the top-5 sentence list based on the students' recommendations.

4.1.2 Comparison between the automatically generated summary and the abstract of the article

We considered the abstract of the article as a reference model given by the document authors and we qualitatively compared the abstract with the automatically generated summary, i.e., research issue (ii). We recall that, prior to text summarization, the abstract content was excluded from the original document. Hence, in general, the abstract and the automatically generated summary have no sentences in common. The paper abstract is reported below.

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively

Table 2: Comparison between automatically generated and golden summaries.

Sum. Rank		Prefs.	Pref. Rank
ItemSum top-5 sentences			
1	This ranking, called PageRank, helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web.	19	1
2	Indeed, many of the web search engines have used backlink count as a way to try to bias their databases in favor of higher quality or more important pages.	19	1
3	In this paper, we take advantage of the link structure of the Web to produce a global importance ranking of every web page.	13	3
4	The intuition behind PageRank is that it uses information which is external to the Web pages themselves, their backlinks, which provide a kind of peer review.	9	7
5	Although there is already a large literature on academic citation analysis, there are a number of significant differences between web pages and academic publications.	2	14
Other article sentences			
7	In order to measure the relative importance of web pages, we propose PageRank, a method for computing a ranking for every web page based on the graph of the web.	14	2
-	The World Wide Web creates many new challenges for information retrieval.	12	4
8	PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure.	11	5
6	Indeed, many of the web search engines have used backlink count as a way to try to bias their databases in favor of higher quality or more important pages.	10	6
-	Using PageRank, we are able to order search results so that more important and central Web pages are given preference.	8	8
-	In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.	7	9
-	However, the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and link text.	7	9
-	Unlike academic papers which are scrupulously reviewed, web pages proliferate free of quality control or publishing costs.	7	9
-	PageRank provides a more sophisticated method for doing citation counting.	7	9
-	The reason that PageRank is interesting is that there are many cases where simple citation counting does not correspond to our common sense notion of importance.	7	9

measuring the human interest and attention devoted to them. We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

The first two sentences introduce the context of analysis. Similar information is given by the top ranked sentence selected by the summarizer: *This ranking, called PageRank, helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web.* The third sentence of the abstract goes in detail of the PageRank algorithm. Similar information can be figured out from the second and third sentences selected by the summarizer: *In this paper, we take advantage of the link structure of the Web to produce a global importance ranking of every web page. The intuition behind PageRank is that it uses information which is external to the Web pages themselves, their backlinks, which provide a kind of peer review.* The other sentences of the abstract summarize the experimental results. This information is partly missing in the automatically generated summary, because the sentences of the experimental part contain few salient keywords and many technical details and punctual statements, which are deemed as less informative by the algorithm. As discussed in Section 4.2, integrating user annotations would be particularly helpful for driving sentence selection not only based on simple word frequency counts but also based on the actual user preferences.

4.1.3 Impact of itemset-based model and minimum support threshold on the summary length

The baseline ItemSum algorithm can generate summaries of arbitrary length by varying the configuration setting of the itemset-based sentence selection strategy Baralis et al. [2012]. Specifically, to adapt the length of the summaries to their needs, learners can (i) change the number of extracted itemsets, by properly setting parameter K, or (ii) vary the minimum support threshold *minsup*, i.e., the minimum frequency of occurrence of the mined itemsets.

We experimentally analyzed the impact of parameter K (i.e., the model size) and of the minimum support threshold *minsup* on the number of selected sentences. The model size indicates the number of itemsets included in the itemset-based model, i.e., the number of combinations of terms that will be considered during the sentence filtering processes. By setting relatively low model size values (i.e., from 3 to 6) only the first two summary sentences are selected (see Table 2). Hence, only the key ideas behind the paper are reported in the summary. By varying the size between 7 and 13 medium-size summaries (i.e., from 3 to 8 sentences) are generated, while further increasing the model size results in a large number of selected sentences (e.g., 11 sentences with model size equal to 15) thus the conciseness of the summaries is significantly reduced.

By enforcing a minimum support threshold during the itemset mining process, the combinations of terms that rarely occur in the analyzed document are discarded. When relatively low support thresholds (e.g., 0.1%) are

enforced, a huge number of combinations is considered thus entropy-based itemset selection becomes practically unfeasible (i.e., the algorithm takes more than 4 hours to terminate). On the other hand, by setting relatively high support thresholds (e.g., 10%) potentially informative itemsets are early discarded because they do not satisfy the support threshold. Therefore, although the most appropriate support threshold value to set depends on the analyzed data distribution, thresholds in the range between 6% and 8% are advisable while coping with documents of this category because they produce fairly high-quality models in few seconds.

4.2 Summary evaluation on benchmark collections

This section addresses the experimental evaluation of the proposed approach on benchmark document collections, which have commonly been used to evaluate the performance of general-purpose summarizers (research issues (iii) and (iv)). With the goal of fostering progresses in automatic text summarization, since 2001 the organizers of the Document Understanding Conferences Document Understanding Conference [2004] proposed several summarization contests. Researchers were asked to submit the summaries generated by their newly proposed approaches on benchmark document collections. To quantitatively evaluate summarizer performance, the submitted summaries were compared with those manually written by domain experts by using ad hoc performance evaluators (e.g., the ROUGE toolkit Lin and Hovy [2003]).

We experimentally evaluated the performance of our approach on the DUC'05 SCU-Marked source documents Document Understanding Conference [2004]. Documents are news ranging over different topics. Based on the covered topics, documents are organized in collections. A group of manually-written summaries of the document collections are edited by hand by real end users to produce a set of simple declarative phrases, hereafter denoted as Summary Content Units (SCUs), for each topic. Thanks to the effort of the university of Ottawa, SCUs were mapped to the original document sentences through the Pyramid evaluation system Copeck and Szpakowicz [2005] to generate a set of SCU-enriched document collections, i.e., 27 document collections annotated at the sentence level. Hereafter we will assume document sentences to be annotated if they are enriched with at least one SCU. The textual content of the SCUs annotating each sentence was considered as the content of the sentence annotation. More than 10% of the document sentences were annotated. SCUs are characterized by a weight indicating the pertinence of a SCU to the given topic. SCU weights were measured as the number of sentences in the reference summaries supporting the identification of that particular SCU. In the DUC'05 documents, SCU weights range from 1 to 7. Approximately 35% have minimal weight (1), 30% of the SCUs have maximal weight (7), whereas the intermediate weights (from 2 to 6) are have similar frequency counts and cover altogether approximately 35% of the SCUs. SCU weights are roughly distributed equally across all documents. Hereafter, we considered the SCU weights (normalized between zero and one) as a measure of the skill level of the user generating the SCU, because expert users are more likely to annotate sentence with pertinent annotations than non-expert ones.

4.2.1 Quantitative evaluation of different ranking strategies

The goal of this section is to quantitatively evaluate the usefulness of driving the summarization process by exploiting highlights, annotations, and user skill levels, beyond the original document content (i.e., research issue (iii)) on the DUC'05 collections. As previously done in Chuang and Yang [2000], to quantitatively evaluate the accuracy of the summarization process, we performed a leave-one-out cross-validation. More specifically, for each topic we summarized all the documents except for one and we compared the achieved summary with the remaining (not yet considered) one, which was selected as golden summary. The process is iterated until all the possible combinations of golden summaries and input documents are considered. Golden and automatically generated summaries were compared using the ROUGE toolkit Lin and Hovy [2003], which was adopted as official TAC'11 and DUC'04 tool for performance evaluation[2]. For each summarizer we computed the average performance results, in terms of precision (P), Recall (R), and F1-measure (F1) for both the ROUGE-2 and ROUGE-SU4 evaluation scores. Since we specifically cope with documents ranging over the same topic, the assumption that a document is a representative summary of all the other documents in the collection is acceptable.

We tested the performance of the newly proposed summarizer with different ranking functions on the SCU-enriched document collections and we compared them with that of the benchmark ItemSum summarizer Baralis et al. [2012], which disregards highlights, annotations, and skill levels during the summarization process. Furthermore, we tested also two largely used opensource summarizers, i.e., OTS Rotem [2011] and TexLexAn TexLexAn [2011]. To test the summarizer with *Baseline ranking*, sentences annotated with at least one SCU were labeled as *highlighted*, whereas the SCU content and the user skill levels were disregarded. To test the summarizer with *Content-based ranking*, for each document sentence we considered the corresponding SCU content (if any) and not only the presence of an annotation (highlighted or not). Finally, to test the summarizer with *Skill-based ranking* we considered not only SCUs but also the skill levels of the users. When not otherwise

[2]We used the command: ROUGE-1.5.5.pl -e data -x -m -2 4 -u -c 95 -r 1000 -n 4 -f A -p 0.5 -t 0 -d -a

Table 3: DUC'05 SCU-market collections. Statistically relevant differences in the comparison between *Skill-based ranking* and the other approaches are starred. Best results in terms of recall, precision, and F1-measure are written in boldface.

Summarizer	ROUGE toolkit					
	ROUGE-2			ROUGE-SU4		
	R	Pr	F1	R	Pr	F1
Skill-based	**0.061**	**0.066**	**0.063**	**0.131**	**0.127**	**0.129**
Content-based	**0.061**	0.054*	0.057*	0.115	0.108	0.110
Baseline	0.052*	0.059	0.055*	0.109	0.106*	0.107*
ItemSum	0.050*	0.044*	0.047*	0.097*	0.093*	0.095*
OTS	0.040*	0.047*	0.043*	0.092*	0.094*	0.093*
TexLexAn	0.024*	0.030*	0.027*	0.049*	0.058*	0.054*

specified, for the summarizer we considered the following standard configuration: model size K=5, minimum support threshold minsup=7%, and α=0.7. More details on the impact of these parameters on the algorithm performance are given in Sections 4.1.3 and 4.2.3.

Table 3 summarizes the achieved ROUGE scores. The statistical significance of the pairwise performance differences was evaluated for all the considered datasets and measures by using the paired t-test at 95% significance level.*Skill-based ranking* performed significantly better than all the other approaches in terms of both ROUGE-2 and ROUGE-4 F1-measure (i.e., the harmonic average between precision and recall). Furthermore, *Content-based* and *Baseline rankings* placed second and third, respectively, and they performed significantly better than ItemSum (original version), OTS and TexLenAn in terms of both ROUGE-2 and ROUGE-SU4.

Based on the achieved results, we can conclude that driving document summarization with additional information is more effective than considering only the original document content. Considering highlights and/or annotation content guarantees summarization performance superior to traditional summarizers exclusively relying on the input document collection. However, since annotations may be low-quality or redundant, in some cases the resulting summaries could be biased. To overcome this issue, integrating user skill levels beyond annotations allows selectively weighting the importance of the annotation content based on the level of expertise of the corresponding user. For this reason, the combination of annotations and user skill levels achieved best summarization performance.

4.2.2 Impact of annotations and skills on summary quality

We performed a qualitative comparison between the summaries generated from benchmark documents by considering (i) only the highlights (i.e., sentence annotated or not), (ii) the highlights plus the annotation content, and (iii) the highlights, the annotation content, and the user skill levels. Table 3 reports the summaries consisting of top-3 ranked sentences[3] and generated with the three different ranking functions on a representative document collection ranging over the following topic: *the privatization program of Argentina in 2005*. For each sentence, the corresponding rank value is also reported. Furthermore, for each ranking strategy we also reported the ranks achieved by a selection of non-top ranked sentences (labeled as *Other sentences*).

The top ranked sentence according to the *Skill-based* ranking (all skills) contains the key information behind the news. It was annotated by highly skilled users with fruitful comments, e.g., *Commercial relations improved steadily between the UK and Argentina* with maximal user skill level (1). However, it ranked only third in the summaries generated with *Content-based* rankings, while it is not selected at all by *Baseline*. Hence, user annotations and skills seem to be helpful for boosting relevant information in the summaries that otherwise might go missing or might achieve relatively low rankings in the summary. On the other hand, when their content is on topic, the annotations made by lowly skilled users appear to be relevant as well for summarization purposes. For example, the top ranked sentence in the summary generated by *Content-based* ranking have two pertinent annotations, associated with fairly low skill level (0.43), describing the status of the negotiation between Argentina and the U.K.: *Discussions between British and Argentina companies on sharing South Atlantic oil resources* and *Joint exploration was being discussed*.

When the top-q sentences (with q between 2 and 4) are considered as output summary, the overlapping degrees, in terms of number of sentences in common between the summaries generated by *Skill-based* and those produced by *Baseline* and *Content-based*, are approximately 66%, because of the enrichment of different data sources (i.e., annotations, skills). Instead, while considering the total number of extracted sentences, the overlapping degrees are approximately 75%. On the same documents we also tried to generate multiple summaries tailored to users with different level of expertise. To this aim, we clustered annotations based on the corresponding skill level and then perform multiple summarization runs driven by different annotations clusters. Table 3 reports also the summaries generated with the *Skill-based* ranking by considering only the annotations

[3] *The summary length approximately corresponds to those of the summaries generated by the domain experts from DUC'05 collections*

Table 4: Comparison between summaries of benchmark documents.

Baseline ranking summary	**Rank**
ARGENTINA is seeking UK expertise to help with the next stages of its ambitious privatisation programme.	0.782
'We are very interested in the British experience of privatisation' he said in an interview with the Financial Times.	0.060
The first phase of Argentine privatisation was criticised for its haste and dogged by rumours of corruption	0.060
Other sentences	
Mr Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions, including Baring Brothers, the merchant bank, and ICI, in London yesterday, to discuss the privatisation programme and encourage UK investment in Argentina.	0.058
A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month.	0.050
Content-based ranking	**Rank**
A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month.	0.060
ARGENTINA is seeking UK expertise to help with the next stages of its ambitious privatisation programme	0.059
Mr Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions, including Baring Brothers, the merchant bank, and ICI, in London yesterday, to discuss the privatisation programme and encourage UK investment in Argentina.	0.046
Other sentences	
'We are very interested in the British experience of privatisation' he said in an interview with the Financial Times.	0.042
The first phase of Argentine privatisation was criticised for its haste and dogged by rumours of corruption.	0.040
Skill-based ranking (all skills)	**Rank**
Mr Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions, including Baring Brothers, the merchant bank, and ICI, in London yesterday, to discuss the privatisation programme and encourage UK investment in Argentina.	0.084
ARGENTINA is seeking UK expertise to help with the next stages of its ambitious privatisation programme.	0.071
The first phase of Argentine privatisation was criticised for its haste and dogged by rumours of corruption.	0.053
Other sentences	
'We are very interested in the British experience of privatisation' he said in an interview with the Financial Times.	0.049
A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month.	0.049
Skill-based ranking (highest skill)	**Rank**
Mr Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions, including Baring Brothers, the merchant bank, and ICI, in London yesterday, to discuss the privatisation programme and encourage UK investment in Argentina.	0.088
A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month.	0.069
The first phase of Argentine privatisation was criticised for its haste and dogged by rumours of corruption.	0.068
Other sentences	
ARGENTINA is seeking UK expertise to help with the next stages of its ambitious privatisation programme.	0.067
'We are very interested in the British experience of privatisation' he said in an interview with the Financial Times.	0.067
Skill-based ranking (lowest skill)	**Rank**
Mr Guido di Tella, Argentine foreign minister, met representatives of British companies and financial institutions, including Baring Brothers, the merchant bank, and ICI, in London yesterday, to discuss the privatisation programme and encourage UK investment in Argentina.	0.094
'We are very interested in the British experience of privatisation' he said in an interview with the Financial Times.	0.083
ARGENTINA is seeking UK expertise to help with the next stages of its ambitious privatisation programme.	0.072
Other sentences	
The first phase of Argentine privatisation was criticised for its haste and dogged by rumours of corruption.	0.062
A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month.	0.058

with highest skill level (1) and lowest skill level (0.14), respectively. In the summaries tailored to users with high skill level more technical content appears and ranked first. For example, the summary driven by lowly skilled user annotations mentions the Falklands war only the last sentence, whereas the other covers this aspect in the second sentence by implicitly referring to the Falklands war. Unskilled users may be unaware of past events thus they could not understand the connection between the current dispute and the historical facts.

4.2.3 Analysis of the algorithm parameters

The newly proposed summarizer allow users to choose to what extent annotations affect the sentence relevance score and, thus, their likelihood to be included in the summary. In the computation of the sentence ranking $\text{Rank}(s_i^j)$, parameter α varies between zero and one. The higher α we set, the more important user annotations are with respect to the document content (for more details see Section 3.3). We empirically analyzed the impact of parameter α on the summarizer performance. Figure 2 plots the average ROUGE-SU4 Precision, Recall, and F1-measure achieved on the DUC'05 SCU-marked collections by varying parameter α between zero and one. Setting intermediate values (i.e., between 0.4 and 0.7) yields fairly good performance, whereas setting relatively

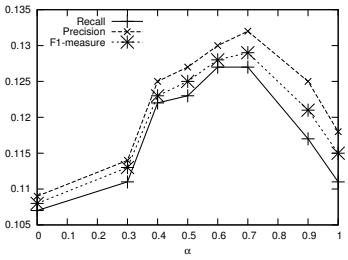

Figure 2: Impact of parameter α on the ROUGE-SU4 F1-measure.

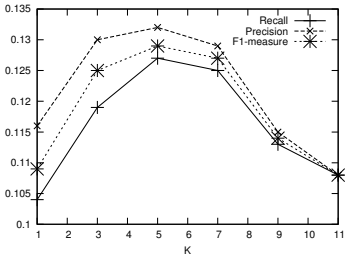

Figure 3: Impact of parameter K on the ROUGE-SU4 F1-measure.

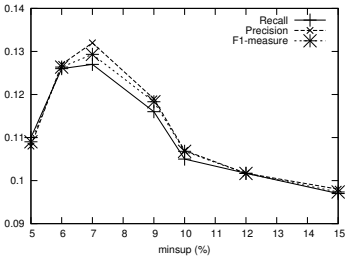

Figure 4: Impact of parameter *minsup* on the ROUGE-SU4 F1-measure.

low/high values results in unsatisfactory results because the influence of one data facet (i.e., the documents or the annotations) is almost ignored.

In Figures 3 and 4 we plotted the variation of the average ROUGE-SU4 Precision, Recall, and F1-measure achieved on the DUC'05 SCU-marked collections by varying parameters K and *minsup*, respectively. The curves achieved on the real documents analyzed in the crowd-sourcing experience are similar to those reported in this section.

Using the content-based ranking function instead of the skill-based ranking curve trends are similar for α values between 0.4 and 0.7, whereas they are slightly more flattened for high α values, because unreliable annotations made by skilled users may significantly degrade the quality of the generated summaries. By setting high α values (>0.7) or low K values (<5) Recall is slightly lower than Precision because a few interesting correlations among document terms are ignored. In all other cases, Recall and Precision values are pretty close to their harmonic average (F1-measure).

5 Conclusions and future work

The paper investigates the application of itemset-based summarization in support of e-learning activities. It proposes three new summarizers that exploit text highlights, annotations, and user skills, respectively, to drive the summarization process. According to the type of available information different sentence ranking functions are proposed. When the user skill levels are available, the system allows generating also multiple summaries tailored to users with different level of expertise. To demonstrate the actionability of the proposed approach in a real e-learning context, a crowd-sourcing experience of evaluation of the generated summaries was conducted. Furthermore, the impact of annotations and user skill levels on summarizer performance was evaluated on benchmark documents. Future developments of this research will entail (i) summarizing learning documents ranging over different subjects, (ii) studying the impact of the user skill levels on the quality of summaries generated from real teaching materials, (iii) evaluating summaries tailored to different skill levels by involving students with different education levels, and (iv) clustering document sections based on the main topic they

cover and then generating one summary per topic.

References

E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, and S. Shah. Multi-document summarization based on the yago ontology. *Expert Systems with Applications*, 40(17):6976–6984, 2013a.

Elena Baralis, Luca Cagliero, Alessandro Fiori, and Saima Jabeen. Multi-document summarization exploiting frequent itemsets. In *In Proceedings of the ACM Symposium on Applied Computing (SAC 2012)*, 2012.

Elena Baralis, Luca Cagliero, Naeem A. Mahoto, and Alessandro Fiori. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Inf. Sci.*, 249:96–109, 2013b.

Elena Baralis, Luca Cagliero, and Laura Farinetti. Generation and evaluation of summaries of academic teaching materials. In *Proceedings the 39th Annual International Computers, Software and Applications Conference (COMPSAC'15)*, 2015a. doi: 10.1109/COMPSAC.2015.15.

Elena Baralis, Luca Cagliero, Alessandro Fiori, and Paolo Garza. Mwi-sum: A multilingual summarizer based on frequent weighted itemsets. *ACM Trans. Inf. Syst.*, 34(1):5:1–5:35, September 2015b. ISSN 1046-8188. doi: 10.1145/2809786.

Giang Binh Tran. Structured summarization for news events. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 343–348, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2038-2.

Wen-Hsuan Chang, Jie-Chi Yang, and Yu-Chieh Wu. A keyword-based video summarization learning platform with multimodal surrogates. In *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*, pages 37–41, July 2011. doi: 10.1109/ICALT.2011.19.

Wesley T. Chuang and Jihoon Yang. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the 23rd ACM SIGIR conference*, SIGIR '00, pages 152–159, New York, NY, USA, 2000. ACM. ISBN 1-58113-226-3. doi: http://doi.acm.org/10.1145/345508.345566.

John Conroy, Judith Schlesinger, Jeff Kubina, Peter Rankel, and Dianne OLeary. Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics. In *TAC'11: Proceedings of the The 2011 Text Analysis Conference*, 2011.

Terry Copeck and Stan Szpakowicz. Leveraging pyramids. In *DUC'05: Proceedings of the 2005 Document Understanding Conference*, 2005.

Document Understanding Conference. HTL/NAACL workshop on text summarization, 2004.

Mark Dredze, Hanna M. Wallach, Danny Puller, and Fernando Pereira. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 199–206, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-987-6. doi: 10.1145/1378773.1378800.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Bernd Bohnet, Yang Liu, and Shasha Xie. The ICSI/TUD summarization system at TAC 2009. In *Proceedings of the Text Analysis Conference*, TAC '09, Gaithersburg, MD (USA), 2009.

Oskar Gross, Antoine Doucet, and Hannu Toivonen. Document summarization based on word associations. In *Proceedings of the 37th ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1023–1026, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609500.

M.S. Hossain, M. Masud, A.A. Alelaiwi, and A. Alghamdhi. Aco-based media content adaptation for e-learning environments. In *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, pages 118–123, May 2014. doi: 10.1109/CIVEMSA.2014.6841449.

Jiri Hynek and Karel Jezek. Practical approach to automatic text summarization. In *ELPUB*, 2003.

M.-B. Ibanez, A. Di-Serio, and C. Delgado-Kloos. Gamification for engaging computer science students in learning activities: A case study. *Learning Technologies, IEEE Transactions on*, 7(3):291–301, July 2014. ISSN 1939-1382.

Szymon Jaroszewicz and Dan A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 178–186, 2004.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 71–80, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526720.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 71–78, 2003.

Michael Mampaey, Nikolaj Tatti, and Jilles Vreeken. Tell me what I need to know: Succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

Joi L. Moore, Camille Dickson-Deane, and Krista Galyen. e-learning, online learning, and distance learning environments: Are they the same? *The Internet and Higher Education*, 14(2):129 – 135, 2011. ISSN 1096-7516. doi: http://dx.doi.org/10.1016/j.iheduc.2010.10.001.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. Technical Report 1999-66, November .

Mohsen Pourvali and Mohammad Saniee Abadeh. Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base. *CoRR*, abs/1203.3586, 2012.

N. Rotem. Open text summarizer (ots). retrieved from http://libots.sourceforge.net/ in july 2011, 2011.

S. Saraswathi, M. Hemamalini, S. Janani, and V. Priyadharshini. Multi-document text summarization in e-learning system for operating system domain. In Ajith Abraham, JaimeLloret Mauri, JohnF. Buford, Junichi Suzuki, and SabuM. Thampi, editors, *Advances in Computing and Communications*, volume 193 of *Communications in Computer and Information Science*, pages 175–186. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-22725-7. doi: 10.1007/978-3-642-22726-4$_$19.

Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, and Vanni Zavarella. Jrc's participation at tac 2011: Guided and multilingual summarization tasks. In *TAC'11: Proceedings of the The 2011 Text Analysis Conference*, 2011.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN 0321321367.

TexLexAn. Texlexan: An open-source text summarizer. retrieved from http://texlexan.sourceforge.net/ in july 2011, 2011.

K.S. Thakkar, R.V. Dharaskar, and M.B. Chandak. Graph-based algorithms for text summarization. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on*, pages 516 –519, nov. 2010. doi: 10.1109/ICETET.2010.104.

Dingding Wang and Tao Li. Document update summarization using incremental hierarchical clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 279–288, 2010.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. Integrating document clustering and multidocument summarization. *ACM Trans. Knowl. Discov. Data*, 5:14:1–14:26, August 2011. ISSN 1556-4681. doi: http://doi.acm.org/10.1145/1993077.1993078.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 745–754, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010016.

Guangbing Yang, Dunwei Wen, Kinshuk, Nian-Shing Chen, and E. Sutinen. Personalized text content summarizer for mobile learning: An automatic text summarization system with relevance based language model. In *Technology for Education (T4E), 2012 IEEE Fourth International Conference on*, pages 90–97, July 2012. doi: 10.1109/T4E.2012.23.

Ahmad Issa Saleem Al Zoubib and Mohd Zalisham Jali. An integrated success adoption model for examining e-learning among adult workers in jordan. In *Computer and Information Sciences (ICCOINS), 2014 International Conference on*, pages 1–4, June 2014. doi: 10.1109/ICCOINS.2014.6868829.