


A statistical approach for modeling inter-document semantic relationships in digital libraries

Jeyavaishnavi Muralikumar¹ · Sri Ananda Seelan¹ ·
Narendranath Vijayakumar¹ ·
Vidhya Balasubramanian¹ 

Received: 7 February 2016 / Revised: 1 July 2016 / Accepted: 3 July 2016
© Springer Science+Business Media New York 2016

Abstract E-Learning repositories and digital libraries are fast becoming important sources for gathering information and learning material. Such systems must therefore provide services to support the learning needs of their users. When a retrieval system shows how its documents relate to each other semantically, a user gets the liberty to choose from different material, and direct his/her study in a focused manner. This calls for a model that identifies types of document relationships, that need to address different aspects of learning. This article defines three such types and a unique statistical model that can automatically identify them in technical/scientific documents. The model defines measures to quantify the degree of relatedness based on distinct statistical patterns exhibited by the common terms in a pair of documents. This approach does not strictly require a knowledge base or hypertext for identifying the characteristic relationship between two documents. Such a statistical model can be extended to build further relatedness types and can be used alongside various other techniques in digital library recommendation engines. Our experiments over a large number of technical documents show that our techniques effectively extract the different types of relationships between documents.

Keywords Relatedness · Information retrieval · Digital libraries · Statistical modeling

1 Introduction

Automated guidance for users of digital libraries is provided through suggestions based on the content similarity of the documents, citations, or by matching other meta-data like authors, keywords, topics, etc. However, retrieval over such libraries must also provide

✉ Vidhya Balasubramanian
b_vidhya@cb.amrita.edu

¹ Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

information in a way that the learning objectives of a person seeking information are met, and ensure that the browsing experience remains focused. This applies especially to scientific documents, such as in the case of a person doing a scientific literature survey.

Several such learning objectives that must be addressed have been proposed. With respect to scientific literature, the model proposed by McCormak and Yager (Frantzi et al. 2000), outlines the different aspects of learning like ‘knowing’, ‘exploring’ and ‘using’. Services that can support automatic suggestions to enable the user experience these aspects are desirable in a digital library.

Keeping in mind these aspects of scientific learning, we first identify different aspects or dimensions in which documents are related so that it will be helpful for a learner. We consider three factors that are important in helping a learner browse through a repository of technical/engineering documents. When exploring a new set of concepts, the user might not have sufficient knowledge to,

- decide what documents to look at next in the sequence
- know where to look for a more detailed explanation of a concept, when the concept is insufficiently described in the current document
- see where the concept may be applied if this information is not present in the current document

An automated system modeled to provide such document suggestions, that can enable the learner overcome these kinds of difficulties is desirable. Such a system must be able to identify the nature of connection or relationship that occurs between the content of documents, and present appropriate suggestions at each stage of the learning process. While there are plenty of models for identification of concept and semantic relatedness, definition of document relationships has been less extensively studied (Khoo and Na 2006). This gives rise to the need for a novel relatedness model for identifying semantic relationships between documents from a learner’s perspective. Our work focuses on relationships that capture the above three requirements. Specifically we are interested in determining if a document d_i (or a part of it) represents a ‘*continuation*’, ‘*elaboration*’ or ‘*application*’ of a document d_j (or a part of it).

There are two approaches for identifying these relationships between two documents. One is to use a knowledge based approach, where the concepts and their relationships are modeled, and the other is to use a statistical approach. In the former, known relationships between concepts in the documents have to be used to infer the underlying relationship between the documents. Knowledge based approaches can be used to directly infer concept level relationships; however, techniques have to be developed to infer the above relationships from concept level relationships. For instance, consider a document A that is about *stacks*, that briefly mentions *linked lists* from an implementation perspective. While there is a conceptual relationship between *stacks* and *linked lists*, the presence of the concept *linked lists* in another document, say B , does not necessarily make it the right document for giving a detailed insight into *linked lists*, unless we can be certain that *linked lists* is indeed the major topic in document B .

In addition, knowledge sources are only available for some domains and for many other domains are not comprehensively defined or are yet to be generated. Therefore, a technique to automatically identify the dominating relationship(s) between two documents and to quantify the strength of the relationships without much domain knowledge is essential. It is desirable to infer these relationships from the documents themselves, and hence we use that

as a start point for addressing this problem. In this article, we intend to present a comprehensive approach that identifies such requirements without relying on external knowledge. In future it can be extended to include semantic knowledge derived from external sources.

In order to identify the inter-document relationships, we use an approach that analyzes the statistical patterns within documents and uses these to infer their semantic relationship. We approach this problem of identifying and characterizing document relationships, with a premise that the semantic structure of a document is brought out by the occurrence patterns of terms and their context of usage in the document. Quantification of these aspects can be done using statistical measures. For example, measures like C-Value and TF-IDF (Frantzi and Ananiadou 1996; Balagopalan et al. 2012) characterize the degree of importance of a term in a document.

We characterize the relationships between two documents using such features. Such a characterization constitutes our relatedness model to identify and quantify the different kinds of relationships between documents. The final goal of this relatedness model is to provide scores based on these measures that indicate the strength of such relationships between documents. Such scores can be used to provide suggestions of related work and guidance to users of digital libraries, along with an indication of how the suggested work is related. Such functionality would further the development of recommendation systems and innovative visualizations, that can aid the users of a digital library. Our specific contributions in this paper are the following:

- Defining types of document relatedness that would serve learning needs of consumers of scientific/technical content
- Identifying patterns between documents that characterize different relationships, based on extensive analysis of sample document data sets
- Development of measures to quantify different types of relationships using the identified features
- Extensive experimentation and evaluation as that validates the above

The rest of this article is organized as follows: Section 2 outlines the related work and the following section introduces the three relationships and describes the approach taken to identifying characteristics of these relationships, that can contribute to detecting/quantifying them. In Section 4 we give formal definitions of these relationships, and also introduce measures to quantify the strength of these relationships for a particular instance. These are validated in Section 5 by describing extensive experimental results. Finally, we conclude in Section 6.

2 Related work

Relatedness has been widely studied in the context of concepts, paragraphs, terms and documents, and there have been many measures that solve the problem of quantifying semantic similarity and semantic relatedness. The most fundamental relation that can occur between documents is the similarity relation. Cosine similarity and Jaccard similarity coefficients are examples of statistical measures that can be used to find the similarity between documents or text fragments. The cosine measure is fundamental, as it is used in other techniques for estimating similarity, such as those that rely on external knowledge, as well as those that don't (i.e., LSA).

Similarity measures that involve computing semantic similarity or relatedness through analysis of an external knowledge base sometimes use a reference corpora. The Normalized Relevance Distance (Schaefer et al. 2014) is one such recent corpus-based measure, built on the Normalized Compression Distance. Otherwise, they typically use a graph such as WordNet (Agirre et al. 2009; Capelle et al. 2013) or Wikipedia, or the World Wide Web. A well established method is the Explicit Semantic Analysis (ESA). Gabrilovich et al in Gabrilovich and Markovitch (2007) use ESA with the Wikipedia document corpus as the knowledge base and compute the relatedness measure between documents using cosine similarity. Other approaches based on Wikipedia hyperlinks Gouws (2010), Strube and Ponzetto (2006), Turdakov and Velikhov (2008). While using Wikipedia has its advantages, as argued by Denning et al. (2005), using such a knowledge base has its own risks like uncertain expertise of contributors.

In contrast to ESA, Latent Semantic Analysis (LSA) (Foltz et al. 1998) is used in Natural Language Processing to find the degree of similarity between two pieces of text without an external knowledge base. Here the similarity is found by using the cosine index method. Probabilistic Latent Semantic Analysis (Hofmann 1999), another method that evolved from LSA uses a probabilistic model to find similarity between documents. Both LSA and probabilistic LSA account for issues such as synonymy and polysemy but are used to determine only similarity or association relations whereas a more comprehensive set of relations is required in a digital library scenario.

User interest-based approaches like collaborative filtering involve user profiling in recommendation systems (Hopfgartner 2010; Lai et al. 2013; Huynh et al. 2012) and may account for relations beyond similarity. Though such techniques identify connections between different documents, the problem of quantifying the different types of context-based relatedness between them remains unsolved.

Citation links have been widely used to capture document relationships ((Zarrinkalam and Kahani M. 2012) is one example), but do not always guarantee a semantic relation (Bean and Green 2001). Also, methods involving hypertext are limited to structured text documents and cannot be used for unstructured text, video/audio transcripts or documents such as lectures which do not have citations, links, headings or sub-headings.

Similarity measures have also been used for the classification and clustering of documents. It has been stated that the fundamental measures such as the cosine similarity, Pearson coefficient and KLD are effective at identifying similar documents. Works like Huang (2008), Chalmers and Chitson (1992) and Andrews et al. (2001) describe clustering of similar documents. Researchers have defined similarity measures that are based on topic/concept representations of the documents as discussed in Rafi and Shaikh (2013) and Huang et al. (2012). Works like Wan and Peng (2005) and Wan (2007) define document similarity measures that are based on similarities between different segments of the document, thus taking term distribution into account.

Much of the above work focuses on the generic similarity relation between documents. Gonzalez-Agirre et al. (2015) lists some types of semantic similarity between documents. These works (similar to Aletras et al. (2012)) concentrate on digital libraries with cultural items. However, work on types of relatedness in an educational context for the particular case of academic/educational documents is hard to come across.

In this article, we define a relatedness model that serves the learning needs of users in such a context, which is generic enough to be meta-data independent. We don't directly opt for a graph-based or knowledge-based approach since, it may not always be feasible to analogize a document to a particular concept or a Wikipedia article, not least because said piece of literature may feature multiple topics. Therefore, we go for a statistical model that

uses the context of term usage and other intrinsic term features to identify inter-document relationships.

3 Relatedness modeling

Our goal in this section is to model the different types of relatedness using different statistical measures. The scope of this paper is to meet the needs of a user in a digital library which typically is a heterogeneous collection of educational content such as scholarly articles, video and audio lectures. Our measures should be generic so as to work for all of these types of content. Additionally, digital libraries are often intended for learning or research requirements, and the relationship types between documents must serve such requirements.

Therefore the relatedness types are chosen so that they correlate with the five aspects of teaching and learning science education literature proposed by McCormak and Yager (Frantzi et al. 2000), which are 1) Knowing and Understanding 2) Exploring and Discovering 3) Imagining and Creating 4) Feeling and Valuing and 5) Using and Applying. While there are several models that refer to aspects of teaching and learning, for scientific documents, the Yager model is the most appropriate.

We map three of these categories of learning objectives which can be primarily related to information access, to the potential document relatedness categories. These relatedness categories are *Continuation*, *Application/ Implementation* and *Elaboration*. The other aspects i.e., imagining/creating and feeling/valuing deal with factors beyond the scope of information access in digital libraries and hence are not considered.

A person making a literature survey by browsing a repository of documents tries to “Explore and discover” information through a series of logical sequences or paths. Any such path can lead to discovering information. However, we focus on providing a channelized series of document suggestions, the simplest would be when the information flow is continuous. This type of an exploration process of a learner motivates the presence of the “Continuation” type between two documents, i.e., when there is a natural progression from the content of a document to another.

“Application/ Implementation” directly correlates to the “Using/Applying” category. This relationship helps the learner to know about how a particular concept defined or mentioned in a document is applied or used in another document.

An inquisitive user who needs a comprehensive understanding about the concept that is discussed in a document gets benefited when documents are presented that consist of information which are “Elaboration” of that particular concept. This directly corresponds to the “Knowing and understanding” aspect. While the “Continuation” type helps the user to browse through a set of documents that reflect a single continuous flow of concepts, the latter two types serve more specific purposes which involve presenting a diversified chain of concepts helping the learning process. In short these three types of document relatedness correspondingly answer the following queries of the user:

- “I understand this lecture/material. What document do I look at next?”
- “I have understood the concept described in this lecture/document, but where and how is it actually used?”
- “I need a better understanding of this concept in this document. Where can I find more information on it?”

In the rest of this section, we discuss base patterns that help quantify the strength of these relationships.

3.1 Statistical term features

Two documents are related by the content they share, including common terms or phrases. We use these common terms or phrases to develop a statistical model for identifying the type of relation they may have. To do so, we consider the following properties of such terms:

- The relevance of a term or phrase to each document.
- The distribution or occurrence pattern of a particular term in each document.

The relevance of a term indicates how significant it is, and how much it represents the topic featured in the document. This is important in the identification of the specific relation between the documents. In addition, each relation is defined by using the occurrence patterns of these terms rather than just their occurrence or term frequency. We therefore identify features that quantify either of these properties. These features are the C-Value and TF-IDF for measuring relevance, and D-SPAN for measuring occurrence patterns. They are defined as follows.

C-value We make use of the C-value defined in Frantzi and Ananiadou (1996). This measure quantifies the relevance of a term to the subject of its document. McCormack and Yager (1989) discusses how to use C-value as a context measuring factor for terms present in a document. C-Value looks at both linguistic and statistic features to determine if a term is potentially a keyword or keyphrase. A term has a high C-value if it has high frequency and occurs infrequently as part of longer candidate phrases, which makes it very relevant to the main theme of the document as described in Balagopalan et al. (2012). For a term a , C-Value is defined based on its length $|a|$, frequency $f(a)$ and the set of larger terms that contain a (T_a with length $|T_a|$). The mathematical definition of C-Value is given by (1).

$$C\text{-Value} = \begin{cases} \log_2 |a| \cdot f(a), & \text{if } a \text{ is not nested} \\ \log_2 |a| \cdot (f(a) - (\frac{1}{|T_a|} \cdot \sum_{b \in T_a} f(b))), & \text{otherwise} \end{cases} \quad (1)$$

TF-IDF The TF-IDF is an elementary feature in Information Retrieval and defines the relevance of a term to a document as a product of term frequency and inverse-document frequency.

D-SPAN D-SPAN is a measure of both how widely a term is distributed in a document, and the frequency of occurrence of the term in the document. To measure how well a term a is spread across a document we use the feature “Dispersion”, defined in Balagopalan et al. (2012). The occurrences of a term a are represented as I_a , with each element in I_a being the number of words between successive clusters of the occurrence of a . Each cluster is the set of occurrences of a within k words of each other, where k is experimentally determined. Dispersion, then, is defined as the ratio of the length of I_a , ($|I_a|$) to the variance of I_a ($var(I_a)$).

$$Dispersion(a) = \frac{|I_a|}{var(I_a)} \quad (2)$$

We define D-SPAN as follows:

$$D\text{-SPAN} = \text{wordcount} \times Dispersion \quad (3)$$

$$\text{wordcount} = TF(a) - \sum TF(t) \quad \forall t \text{ in } T_a \quad (4)$$

In this equation, ‘wordcount’ is the term frequency, after eliminating the occurrences of the terms as part of larger n-grams. Terms that occur throughout and often in a document have higher D-SPAN values.

3.2 Identifying patterns in term features

In order to provide a starting point for the definition of our intended measures of the types of relatedness, we made some exploratory visualizations and we observed that studying patterns in D-SPAN and C-Value was helpful. We start with two basic premises:

- Documents featuring similar subjects use the same terms to similar extents, and hence will have common terms with similar D-SPAN values.
- Documents that have related, but not similar subjects would have the same terms, but in different contexts. Hence, they will feature common terms with differing C-Values.

Let us consider two documents D_1 and D_2 , having a collection of t common terms, T . Every i^{th} term in T is present in both D_1 and D_2 and has a D-SPAN and C-Value in both D_1 and D_2 . Consequently, every term in T has a difference or similarity in its D-SPANS or C-Values in the two documents. This difference or similarity would indicate the varying context in which this term is used, and could potentially help identify how D_1 and D_2 are related.

We consider the average C-Value difference and the aggregate D-SPAN similarity for all terms in T , and see if these are indicative of any type of relatedness. For a set of n documents, we would have $n \times n$ pairs of documents. Each pair of documents have t common terms, with $D-SPAN_n(i)$ being the D-SPAN of common term i in document n , and $C-Value_n(i)$ being the C-Value of common term i in document n . Each such document pair, (D_1, D_2) is then characterized by (D_{sim}, C_{diff}) , where:

$$D_{sim} = \sum_{i=1}^t 1 - |D-SPAN_1(i) - D-SPAN_2(i)| \quad (5)$$

$$C_{diff} = \frac{1}{t} \sum_{i=1}^t C-Value_1(i) - C-Value_2(i) \quad (6)$$

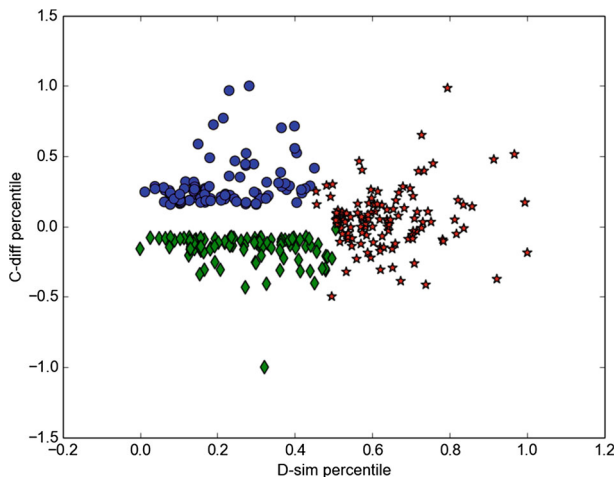


Fig. 1 BIRCH clustering of nxn document pairs

Table 1 Evaluation of clustering result for $n \times n$ document pairs

Balanced F-score	0.59
Normalized Mutual Information	0.6

After eliminating document pairs with feature values below a certain threshold, the remaining D_{sim} and C_{diff} values are min-max normalized and run through an implementation of BIRCH (Zhang et al. 1996), with an unspecified number of clusters. The result for an original data set of 123 documents is shown in Fig. 1. We do not include all $n \times n$ document pairs in our visualizations, since it can hardly be expected that all 123^2 pairs will feature a semantic relation. As a means of reducing the noise, we only included document pairs which had D_{sim} or C_{diff} values higher than the majority of the data points. The filter was determined arbitrarily by making a histogram, and eliminating the densest bins.

For evaluation, we took smaller sample sets out of the $n \times n$ document pairs, and calculated balanced F-score and NMI. Each of the sample sets had data points for one particular document, and had about 15 document pairs. Based on expert estimation, the document pairs were classified as exhibiting Continuation, Elaboration or Application, and F-Score and NMI were calculated between the expert's classification and the clustering result. The average NMI and F-Score values are shown in Table 1, and are promising. The NMI value is an indication that clustering based on these features is informative as to determining the relationship between D_1 and D_2 . However, the moderation of the NMI and the F-Score is also an indication that precision of the classification should be improved, and it is necessary to augment these patterns and to come up with reliable measures.

Consider a document pair D_1 , D_2 with D_1 as the source document, and D_2 being its continuation, application or elaboration. Our particular observations are as follows:

- Document pairs exhibiting continuation have tended to end up in the far right cluster, consisting of data points where aggregate D_{sim} is high
- Document pairs exhibiting elaboration have been observed in the bottom-left cluster, with negative C_{diff} , indicating higher C-Value of common terms in the elaboration document (D_2)
- Document pairs exhibiting application have been commonly observed in the top-left cluster, where C_{diff} is positive, indicating higher C-Value for common terms in the source document (D_1)

For the set of documents tested here, their pairwise relationship was determined using domain experts. This serves as the basis for the above analysis.

The likelihood that a particular relationship R (say, continuation) is characterized by the corresponding pattern (in this case, high D-SPAN similarity among common terms), can be simply estimated by applying the Bayes rule for a random set. Let us consider a set of data points with the above mentioned features. The classical probability of a data point belonging

Table 2 Estimation of Bayesian probabilities that types of relatedness characterized by specific feature score patterns

Type of relatedness	Score pattern	Estimated probability
Continuity	Medium to high D_{sim}	0.75
Elaboration	Negative C_{diff} ; C-Values higher in the elaboration	0.65
Application	Positive C_{diff} ; C-Values lower in the application	0.59

to a particular cluster C_1 can be readily calculated. Similarly, the conditional probability that a data point in C_1 exhibits R can also be calculated. Doing the same for the complementary cases, we can easily estimate the likelihood that the score pattern represented by the C_1 (such as high D_{sim} or low C_{diff}) indicates the relationship. These probabilities, estimated for a random set, can be found in Table 2. This random set had 20 different documents, and 37 document pairs were clustered after points with low D_{sim} and low C_{diff} were eliminated. The relation exhibited by a particular pair, was decided by an expert user. These figures indicate that our relatedness measures may be built on these particular patterns, although further validation needs to be done.

We now extend our observations to more formally define these relations in the next section. We also describe measures to quantify each type of relatedness and provide ranked suggestions to the user, motivated by the identified patterns and some additional observations. Later, we validate these measures by extensive evaluation on their performance on varied data sets.

4 Defining and quantifying types of relatedness

Before we define each type of relatedness, we first state the idea that it is not always the most relevant terms that contribute to the relationship between a pair of documents. It is particularly important to keep this notion in mind when we deal with the types elaboration and application. For instance when looking for elaborations, users are usually looking for elaborations of less prominent topics or concepts featured in the media in question, which we may expect not to be represented by the term in the document with the highest C-Value or term frequency. Therefore, there is a need to formalize different classes of common terms that can play a role in the definition of these relatedness types.

We categorize terms in a document based on their contribution to the content of the document. The role played by a term in a document can be determined based on its context and relevance. From the definitions in Section 3, it can be understood that a term having high C-Value tends to be more contextual to the core content of the document. Similarly, high TF-IDF means the term is very relevant and particular to that document, even if it does not directly relate to its core topic. A good D-SPAN score indicates that the term is used well throughout the document at regular intervals.

- **Topic:** A term can fall under this category if it is the subject of a discussion, lecture or an article, and is characterized by high C-Value, high TF-IDF and high D-SPAN
- **Subtopic:** Subtopic term is also a subject of a discourse, but not the most prominent one. This term is characterized by moderate C-Value, moderate TF-IDF and low or moderate D-SPAN
- **Background:** A background topic term does not define the central theme but is generally a concept that is used to support the discussion of the topic. A background term is characterized by relatively low C-Value and low TF-IDF and low or moderate D-SPAN

For instance, in a lecture on Queues, ‘Queue’ would be a topic term, terms such as ‘circular queue’ and ‘double-ended queue’ would be subtopics if they were talked about for a while, and terms denoting, say, algorithmic complexity or linked lists may be background terms. It is straightforward to train a rule-based classifier to classify terms as topic, subtopic or background. Table 3 shows some more examples of such terms and their classes.

We now proceed to define the three relationships, which we will define on a logical level, as well as based on these categories of terms. We also motivate a definition of measures to

Table 3 Sample feature scores for Topic, Subtopic and Background terms

Term	Title of document	TF-IDF	C-Value	D-SPAN
Topic terms				
Binary search	Binary Search Trees	1	0.92	0.21
Machine Learning	Intro to Machine Learning	1	0.78	0.24
Context free grammar	Grammar	0.78	1	0.45
Quick Sort	Quick Sort	1	1	0.47
Subtopic terms				
Supervised Learning	Intro to Machine Learning	0.23	0.35	0.14
Functional Languages	Syntax	0.23	0.3	0.11
Base address	Pointers	0.43	0.36	0.14
Logical notation	Logical Inference	0.26	0.24	0.13
Background terms				
Increasing order	Quick Sort	0.07	0.08	0.08
Linked list	Trees	0.06	0.06	0.03
Root node	Trees	0.05	0.05	0.03
Left child	Trees	0.09	0.09	0.04

rank documents based on these relationships, by using the patterns mentioned in Section 3.1 to characterize their strength.

4.1 Defining and quantifying continuation

The ‘continuation’ relation provides answers to the query type “What must I study next?”. Here we expect a document D_2 that follows D_1 to lie within the larger subject domain of D_1 , and thereby exhibit some level of similarity in both content and flow of the content. When we say that document D_2 is a continuation of document D_1 , we consider two aspects:

- D_2 can continue the same topic as D_1 , i.e., two documents demonstrate considerable similarity with respect to the usage of topic terms. For instance a lecture on “Hashing and Hash functions” is continued by the lecture on “Universal Hashing”, here topic terms like “hashing”, “hash tables” are spread uniformly and widely in both lectures. D_2 can also continue a subtopic explained towards the end of D_1 . Here again there is considerable similarity in the subtopic terms.
- D_2 and D_1 can talk about the same domain characterized by similarity in background terms, hence indicating a natural progression of topics within the same domain. For instance, a lecture on “Dynamic Programming” can be a potential continuation for a lecture on “Greedy Algorithms”. They would use common concepts such as optimization, analysis, etc which are background topics or common subtopics like “knapsack” if both documents delve into the same examples.

We base our Continuation measure on both the spread of the common terms and the importance of the common terms to the document pair, i.e. the D-SPAN and TF-IDF of common terms. As shown in Section 3.2, similarity in D-SPAN values can be an indicator of the continuation relation. Continuation therefore differs from similarity measures like cosine similarity or LSA. These measures usually only consider the similarity in term frequencies or TF-IDFs, but not the occurrence patterns of the term.

For characterizing continuation, D-SPAN is an important measure. The D-SPAN score is a measure of term usage of a document; Similar D-SPAN scores for a term in two documents means that the term is used to an equal extent and with equivalent scope in both documents. While the spread of common terms is a good indicator of continuation, the relevance of the common terms to the target documents is also an important factor to be considered. Experimentally, we have seen that the effectiveness of the D-SPAN measure is further improved by taking into consideration the TF-IDF value, which has been widely used as a term weighting feature in query-document retrieval.

Therefore we consider the TF-IDF as a direct measure of relevance. Though TF-IDF is not always directly used as a probability function for relevance, it is considered as a measure of the probability of relevance (Wu et al. (2008), Schaefer et al. (2014)). The use of a TF-IDF weight means that the high distribution similarity of more relevant terms indicates a stronger instance of the Continuation relation. This is illustrated with an example in Fig. 2. The combined use of these features means that the D-SPAN similarity of terms with higher relevance defines the Continuation relation. For a strong Continuation relation, the documents must feature high D-SPAN similarity for common terms of varying relevance (as measured by the TF-IDF value of the terms in these documents), whereas documents that do not exhibit this relation are expected to have a high D-SPAN similarity only in common terms of low relevance.

In terms of behavior of topic, subtopic and background terms, documents that are related by Continuation have a high D-SPAN similarity in all these kinds of common terms whereas documents that don't, may only be expected to have a high D-SPAN similarity in the background terms. Taking these factors into account, we define the Continuation measure as:

$$R_c = \frac{1}{n} \sum_{i=1}^n r_i \quad (7)$$

$$r_i = \min_{TF-IDF}(i) \times D_{sim}(i) \quad (8)$$

In the above equation, $D_{sim}(i)$ is the D_{sim} value for the i^{th} common term between the two documents, and $\min_{TF-IDF}(i)$ is the minimum of the TF-IDF values of the i^{th} common term in the two documents. n is the number of common terms.

To boost the effect of terms with high D-SPAN similarity and high minimum TF-IDF, r_i can be further weighted by a constant factor. Minimum required D-SPAN similarity and TF-IDF is empirically determined by analyzing the data sets. We used a 1.3x boost for such

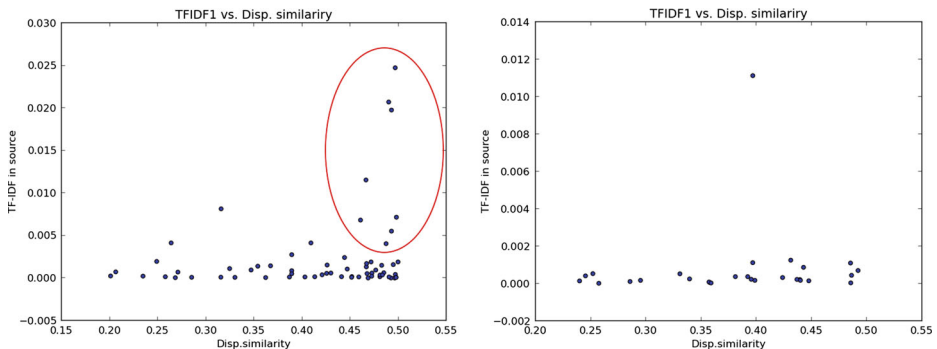


Fig. 2 TF-IDF vs. D-SPAN similarity patterns for “Dictionaries” and “Hash Tables”, which are closely related (left) and “Hash Tables” and “Virtual Memory”, which are related by Application (right)

terms. Using such a boost helps counteract the effect of having a large number of terms of low relevance and low D-SPANs but high D-SPAN similarity, that actually do not contribute to the Continuation relation.

4.2 Defining and quantifying elaboration

This relation helps the user find answers to the type of queries which can be represented as “Where can I find more information about this?”. Usually, the user does not have such a query about the main topic as it would have been explained well in the same document. However, the same may not be the case with the background topics or subtopics. In such cases, elaborations would be those documents in which terms related to background or subtopics are featured more prominently i.e., are topic terms. For example, consider a learner who is studying “Virtual Memory”. There may be references to “Hash Tables” which are commonly used for address mapping. Here, a lecture on “Hashing” would help gain the necessary background knowledge. When a document D_2 serves to provide this background knowledge for some concept that has been briefly covered in document D_1 , D_2 is an elaboration of D_1 .

Since the elaboration relation, by definition, requires concepts to be featured more prominently in D_2 , this implies that common terms are both more relevant and have a higher and wider occurrence in D_2 . We have seen in Section 3.2 that a general trend of higher C-Value in D_2 is evidence for elaboration. This indicates a higher relevance of some terms in D_2 . Since we want these terms to have been explained adequately as to inform and fulfill the ‘knowing and understanding’ learning need of the user, it follows that the D-SPAN of these terms must be greater in D_2 . In other words, the Elaboration relation is defined by the presence of common terms which are background or subtopic terms in D_1 , but have a higher relevance in D_2 .

From our definition of topic, subtopic and background terms, elaboration is characterized by the following features:

- Since a background topic in D_1 is a topic in D_2 , the C-value of the elaborated term will be lower in D_1 . Therefore, we consider common terms with a higher C-Value in D_2 .
- Since D_2 is the elaboration, the common terms are used to a greater extent in it. Therefore D-SPAN of common terms will be greater in D_2 , where they would be topic terms. We count the difference in D-SPAN as a measure of the elaboration relation.

Hence we examine the D-SPAN differences of those common grams which have a higher C-value in D_2 and define the elaboration measure as follows:

$$R_e = \sum_{i=1}^n D-SPAN_2(i) - D-SPAN_1(i) \forall \text{ terms with } C-Value_2(i) - C-Value_1(i) > 0 \quad (9)$$

Equation 9 is meant to be an estimate of the strength of the elaboration relationship. A higher value of R_e means a greater likelihood of the document pair to be exhibiting this relation. That is, the destination document will have a higher D-SPAN value for terms that exhibit higher C-Value in the destination document.

4.3 Defining and quantifying application

An application of a concept helps a learner understand where it is actually used. We can consider documents D_1 and D_2 as being related by application if they satisfy one of the following cases:

- Document D_2 or part of it applies the main concept or a subtopic in a document D_1 . Here topic/subtopic terms in D_1 are usually background terms in D_2 since it is assumed that D_2 is from a different domain or sub domain. For example, “Query Processing” is an application of “Dictionaries”. However a discussion on “Query Processing” will not give much importance to dictionaries. Hence, “Dictionaries” becomes a background topic in “Query Processing”.
- Both D_1 and D_2 share similar background terms, while D_2 applies concepts in D_1 . That is, the application is not explicit, causing the background terms to be present in the applying document. For instance a discussion on Tower of Hanoi applies recursion, thereby also implying an application of a document on stacks. This type of indirect application is not considered in this paper.

Since we expect that the concept being applied is represented by terms that are less prominent in the application, we can expect it to have a lower D-SPAN in D_2 . Coupling this factor with the C-Value difference established in Section 3.2 leads us to a definition of the application measure that is very like the opposite of the previously introduced elaboration measure, even though they aren’t inverses of each other:

$$R_a = \sum_{i=1}^n D-SPAN_1(i) - D-SPAN_2(i) \forall \text{ terms with } C-Value_1(i) - C-Value_2(i) > 0 \quad (10)$$

When the relative C-Value of terms is very high and so is the relative D-SPAN, it indicates that the main topic in D_1 is applied in D_2 . On the other hand, when both the relative C-Value and D-SPAN is moderate, it indicates a subtopic being applied. Based on the equation, the identifying feature of application is the presence of common terms which are topic terms in D_1 , but subtopic terms in D_2 .

Equation 10 is an inverse of the elaboration equation, since it does not consider all aspects of the application relation. It is a preliminary measure, which must be expanded with more features. Currently, we leave this as a direction for future work. Nevertheless, it is rather effective in identifying the application relation, as shown in the next section.

5 Experimentation and results

We have so far discussed our inter-document relatedness model that captures three important aspects of relatedness in scientific document. In this section, we analyze the performance of this model. To evaluate the relatedness model and the measures defined, we develop a prototype system that identifies the relatedness between each pair of documents in a data set.

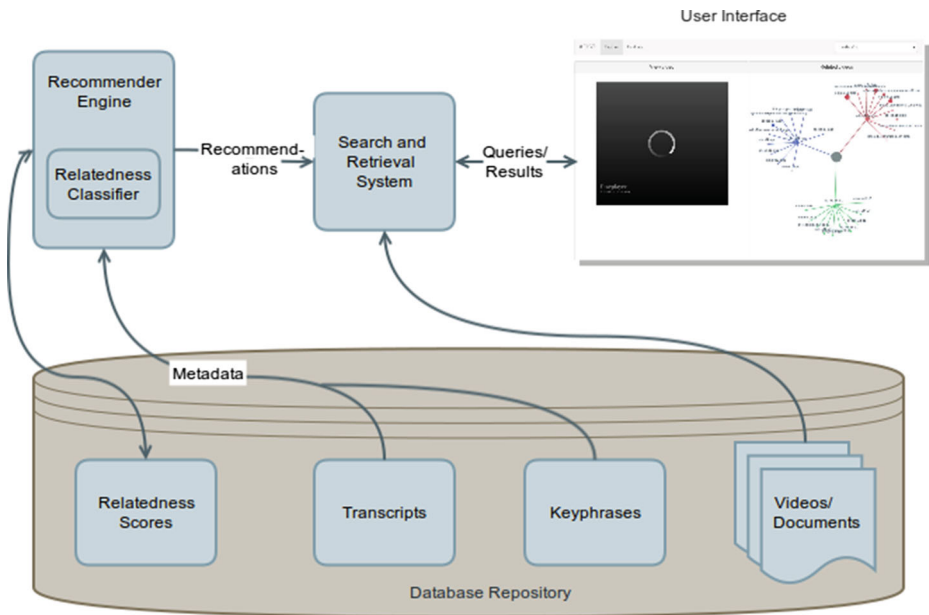


Fig. 3 Architecture of Proposed System

5.1 Prototype system

The architecture of this system is shown in Fig. 3. The prototype system works on a database repository that consists of over 1600 media files, including articles, theses and videos and associated transcripts of classroom lectures. Candidate keyphrases from the documents are extracted after common pre-processing techniques such as stop word removal, stemming, n-gram extraction and POS tagging. Based on the feature vectors representing TF-IDF, C-Value, and D-SPAN scores for the candidate keyphrases (Balagopalan et al. 2012), the document relation scores defined in Section 4 are calculated between each pair of documents. A simple retrieval system is implemented which can retrieve the related documents for a query document, based on these scores. A simple document browser is included, which allows the user to view a selected document, and also displays the results for each category by means of a visualization.

5.2 Datasets and ground truth

One of the primary requirements for evaluating our relatedness model is to choose the right dataset. However, we were not able to find any similar work or datasets, focusing on these specific aspects of relatedness. As mentioned in the related work section, many solutions are available to identify similarity between short snippets of text. Our system, on the other hand, is designed for scientific and educational media, which have considerable length. Therefore, we choose to create datasets for our experiments that take these aspects into account.

We created data sets similar to what would be a typical scientific/ technological document repository that can be found in digital libraries. We specifically gathered a mix of graduate level classroom lecture video transcripts taken from NPTEL (NPTEL 2012) and MIT-OCW (MIT 2012), scientific literature documents, theses, etc. We ensured that the data sets represented several subject domains. They covered multiple document formats and structures, such as classroom lectures and scientific publications. A repository of over 1600 documents was thus created, from which we chose the following data sets:

1. Dataset 1: A set of 120 lecture videos (transcripts), featuring topics such as Database Systems, Data Structures and Machine Learning and Numerical Methods, from NPTEL and MIT-OCW
2. Dataset 2: Contains the transcripts in Dataset 1, as well as a number of scientific publications in the areas of Information Retrieval, Database applications and Geographic Information Systems, totalling over 300 documents.
3. Dataset 3: A set of 1200 lecture transcripts from NPTEL that features topics in Computer Science, Mathematics, Mechanical Engineering, Electrical Engineering, Chemistry etc.
4. Dataset 4: A set of 275 lecture transcripts from MIT-OCW featuring topics in Economics, Physics, Chemistry, Biology and Computer Science

Our prototype system returns a document relatedness score for each pair of documents, and for each category of relatedness. This relatedness score was used to rank suggestions of related documents from the repository. We use two methods to test our measures:

- Create a ground truth for the data set with the help of unbiased experts and compare the obtained results against it
- Gather user feedback from the prototype system and analyze how well the results are in concordance with the intent of users

Ground truth generation For Data sets 1 and 4, a ground truth was generated based on expert judgement. For a chosen subset of documents in these sets, ground truth was determined by generating potential results for sample documents that fall under the different categories of relatedness. While some document pairs are good candidates for continuation and elaboration, they may not be ideal subjects for applications. Hence, the candidate documents were chosen independently for each category based on the data set available and subjects covered. For instance for data set1, the ground truth includes 3 lectures from Data Structures, 2 from Computer Architecture, 2 from computer graphics and 1 each from all other topics. We have therefore ensured that all topics from the data set and different types of lectures (some mathematical, some programming specific, some theoretical) have been chosen. The number of documents was limited so as not to overburden the users with the ground truth generation, which is a time consuming process.

The ground truth was generated by a group of five users for each source document. All of the users have a knowledge of the topic, with an understanding of topics to follow and one of the users was an expert, generally a course handling faculty, or researcher. For each relatedness category the set of documents that hold the particular relationship with the given document was provided by the users after discussion. The users were given time to explore the contents of the lectures in the data set to help with the ground truth creation process.

Table 4 Examples of continuation

Document 1	Document 2	Scaled Cont. Score	Rank
Resistor, Capacitor	Resistors in Series and Parallel	9.8	1
Dictionaries	Hash Tables	7.03	1
Stack	Queue	4.02	2

When there was a disagreement, the expert opinion was sought. In general only when the group arrived at a consensus the document is chosen.

To compare our results against the ground truth, we use **precision** (no. of relevant results/no. of results obtained) and **recall** (no. of relevant results/no. of expected results) metrics.

The second approach where user rating is used to evaluate the results is needed for large data sets, for which it would be difficult to create ground truth. Also, to our knowledge, there aren't any systems or standard data sets for judging these specific relations between documents, which we can compare our system with. Therefore, we also evaluate by user ratings.

We gather user feedback by getting explicit ratings for the recommendation results. A sample from each topic is chosen from the data set for evaluation, and some topics could have more than one sample. Here, for each document D_i the resulting documents categorized under each relatedness type is given to 5 users. Each user gives a rating of 0 to 5, based on the following guidelines:

- 5: indicates that document D2 is definitely related to D1 through the specific relationship.
- 4: indicates that document D2 is strongly related to D1 via the relationship
- 3: indicates that the specific relationship maybe considered but is not definitive
- 2: indicates that this relationship is weak between the two documents
- 1: indicates that the two document definitely do not have the specific relationship but they do have common content
- 0: indicates that the two documents have absolutely nothing in common and there is no conceivable relationship between them

This rating is done individually and privately, and average ratings are presented. The rest of this section details the results of our evaluation.

5.3 Evaluation of continuation

We first evaluate the performance of the continuation relation, based on the results obtained using our data sets. Table 4 shows some sample documents for which the right continuing

Table 5 Evaluation of continuation

	Average Precision	Average Recall
R_c with TF-IDF Weighting	0.83	0.75
R_c without TF-IDF Weighting	0.47	0.72
Cosine Index	0.59	0.76

Table 6 Examples of Elaboration

Document 1	Document 2	Elaboration Score	Rank
Multi-level Indexing	Tree traversal	9.99	1
Computer Graphics: Curves	Cubic Spline Interpolation	4.28	2
Query Processing	Hash Tables	4.72	3

document has been given the high ranks by our measure. These are the document pairs that have been certified by our experts as having the continuation relationship. By ‘Rank’, we mean the ordering according to the Continuation score. The scores have been decimal scaled for the sake of readability, as it does not affect the ranking of the results. The scaled scores are between 0.2 and 22, with a median of 1.15 and the 97th percentile being 4.32. While Dictionaries and Hash Tables have similar patterns for common topic terms, Stacks and Queues have similar patterns in subtopic and background terms. This is a case of our measure working irrespective of the types of terms influencing this relationship, and indicating the effectiveness of our approach. The measure also identifies a good continuing document in other domains like Electronics as shown.

To formally analyze the continuation relation, the continuation scores for document pairs in video lecture transcripts in Data set 1 have been studied. The average precision and recall that were observed is shown in Table 5. As we can see, our measure results in high recall and the use of TF-IDF to weight the D-SPAN similarity of common terms has lead to a better precision. We have shown the results of analyzing the measure without the TF-IDF weighting for comparison.

It can be argued that continuation can be characterized using the similarity relationship. Therefore we evaluate the results obtained when when cosine similarity index (Huang 2008) is applied for the same documents. Average precision of the results obtained when cosine similarity index is applied, is much lower as seen in the Table, while recall is quite similar. The results show that both similarity in occurrence pattern (which is captured by our continuation measure), and the relevance of terms exhibiting these patterns play a role in defining the continuation relationship. Therefore, we are characterizing the Continuation relation by the measure described in Section 4.

5.4 Evaluation of elaboration

Table 7 shows the average precision and recall values for our approach as evaluated with the ground truth. We can see that our approach results in a reasonable precision and recall for this dataset. The elaboration measure uses a filter based on C-Value, i.e., it only takes into account common terms with a higher C-Value in the elaboration document, as described by (9). As proper justification for this choice of feature based filter, we tested the performance of the measure, after replacing the C-Value in (9) with TF-IDF. Even though both

Table 7 Evaluation of Elaboration

	Average Precision	Average Recall
R_e with C-Value as filter	0.53	0.53
R_e with TF-IDF as filter	0.22	0.18

Table 8 Evaluation of Application

	Average Precision	Average Recall
R_d with C-Value as filter	0.48	0.52
R_d with TF-IDF as filter	0.37	0.24

C-Value and TF-IDF seem to bring out the relevance, the degree of contextual relevance is brought out better by C-Value as explained in the definition of these measures. This experiment is a proof for the above premise. The use of C-Value over TF-IDF dramatically improves the results of this measure, demonstrating that the common terms that are relevant to the subject of the document play a major role in this relationship.

Table 6, shows specific examples of our results and the scores using the measure. We can see that the elaboration measure brings out results where a less important concept in Document 1 is prominently featured in Document 2. For instance, the particular lecture on Query processing (Document 1) deals with few techniques that involve hash tables. Though hashing is mentioned often, it is not the most prominent concept in this lecture, making it a background topic in Document 1, while it is the main topic term in Document 2, “Hashing and Hash tables” (Table 7).

5.5 Evaluation of application

From the results shown in Tables 8 and 9, we can observe the utility of the measure in quantifying the application relationship. As in the case of elaboration, we compare the results when C-Value is used over TF-IDF and it is again seen that C-Value proves to be the more effective filter.

From the results we can see that the application measure serves to identify interesting applications. In the examples shown, it has been particularly effective in identifying the usage of the data structures that form the topic of the source document. However, we see from the precision score that it is also prone to false positives, to a greater extent than the elaboration measure. This is because the application measure is rather general and picks up any case where a topic in one document is reduced to a subtopic/background topic in another, while our experts who generated the ground truth actually prefer applications that are novel with respect to the topic being applied. To counteract this effect, identification of additional patterns is necessary, which is a direction for future work.

5.6 Evaluation based on user rating

As mentioned earlier, we evaluate the large data sets by explicit user ratings. Based on the ratings aggregated from 5 independent users, we calculate the average rating given for the top k results for each document of a selected subset of test documents. The ratings were

Table 9 Examples of Application

Document 1	Document 2	Application score	Rank
Dictionaries	Query Processing	24.5	1
Trees	Multilevel Indexing	17.35	1
Trees	Expressions (Programming Languages)	8.92	5

Table 10 Explicit User Rating

	Type of Relation	Average User Rating	Standard deviation	No. samples
Data set1	Continuation	3.34	0.35	50
	Elaboration	3.2	0.66	28
	Application	2.56	0.4	30
Data set2	Continuation	3.34	0.28	25
	Elaboration	3.71	0.21	25
	Application	3.54	0.22	25
Data set3	Continuation	4.07	0.21	30
	Elaboration	3.15	0.30	25
	Application	3.33	0.32	25

given on a scale of 0 to 5. These averages are shown in Table 10. ‘Standard deviation’ is the standard deviation between the users, in terms of rating points, averaged over the total number of document pairs evaluated for that data set and category.

It is seen that the results for “Application” and “Elaboration” depend on the chosen data set. Here, the scores for “Application” are higher for the larger data sets. This is because the number of potential documents that exhibit these types of relatedness is very low in the Data set 1. On the other hand Data set 2 consists of advanced scientific literature, which provide detailed discussions as well as high-level applications. This led to relationships identified between lectures and documents. A few examples are mentioned in Table 11. Data set 3 has lectures spanning a wider range of domains.

We have also collected user ratings for results of Latent Semantic Analysis, with the same samples as the evaluation of Continuation in Data set 2. The initial term-document matrix was of the dimensions 9900×317 . After SVD, rank of the singular value matrix was arbitrarily lowered to 200, and cosine similarity calculated for the document vectors. The rating process was as described earlier. However, the users rated for similarity, as LSA is defined for identifying similarity, and not specific relationships in the educational context. The average rating for the LSA results was 3.04 with a standard deviation of 0.5 rating points. Since it is a similarity measure and judged based on it, we see that LSA needs to be extended to identify other relationships.

Table 11 Results between different document structures

Document 1	Document 2	Type	Rank
ML- Naive Bayes and Maximum margin classifier (lecture)	“Domain-Specific Keyphrase Extraction” (publication; presents method based on the Naive Bayes classifier)	Application	1
Propositional Logic (lecture)	“A Logic-based Approach for Query Refinement in Ontology-based Information Retrieval Systems” (publication)	Application	2
“Generalized Search Trees for Database Systems” (publication)	Trees (lecture)	Elaboration	2

A concept graph depicting the relationship between articles

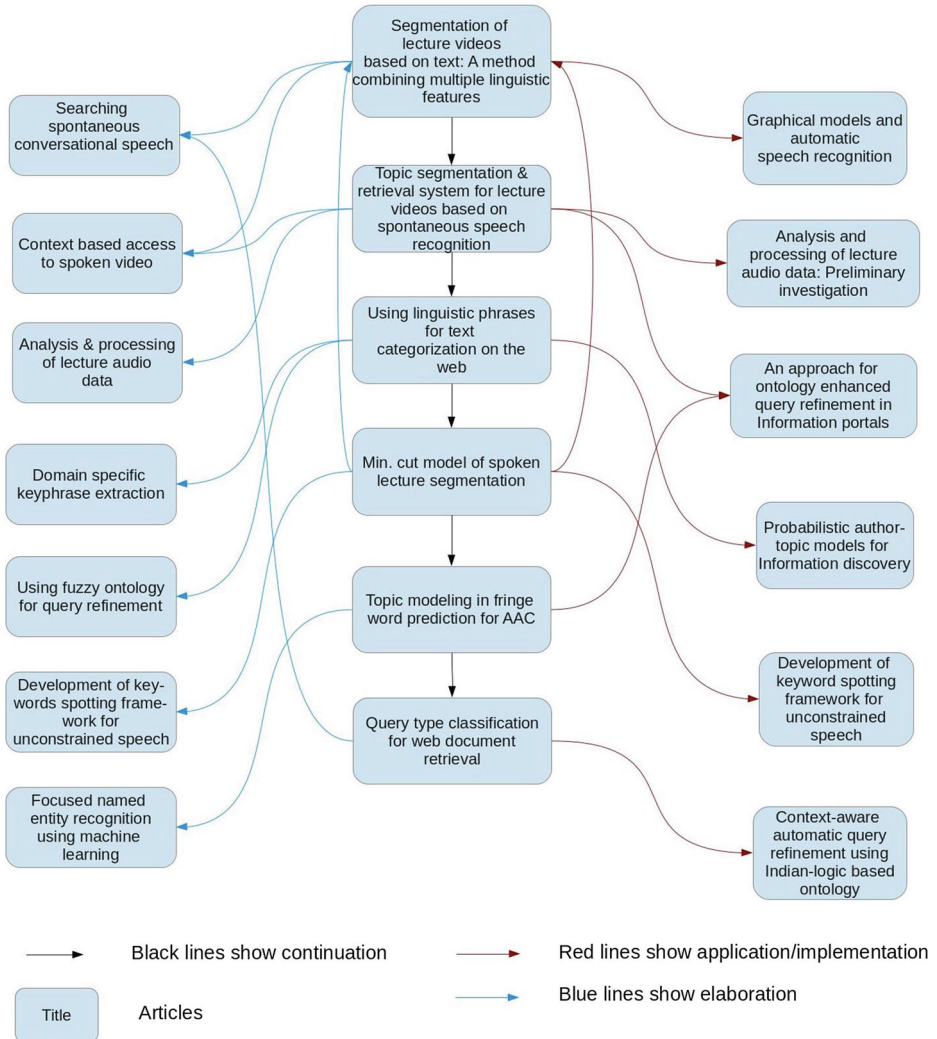


Fig. 4 Document Graph linking Research Articles related to "Segmentation of Videos"

Our experiments have demonstrated that our measures are capable of identifying these relationships between documents and lead to the identification of interesting applications, which could provide useful directions for information search. Figure 4 shows an illustration, which is a manually created document graph based on a sample source document from Dataset 2 - a research article on "Segmentation of Videos". This graph shows recommended continuations, applications and elaborations from the dataset, and represents the extent of the information that can be gleaned from the measures defined in Section 4. This method

of modeling document relatedness presents a more comprehensive view of the available literature, apart from providing more information than a traditional recommendation model, where only suggestions and meta-data are presented. This demonstrates how such a novel relatedness model can help design novel applications for digital libraries.

6 Conclusion

In this article, an approach for identifying semantic relatedness between documents was presented. We described three categories of semantic relatedness: continuation, application and elaboration, and statistical patterns that characterize them. We also defined relatedness measures based on this discussion, and presented our experimental study on their effectiveness. We also demonstrated how such a relatedness model can be effective in generating graphs of related documents. This work represents a starting point for defining different types of semantic relatedness between scientific documents in a digital library. Our methodology serves to identify the three relationships automatically. While the approach works well when the data set is broad, document characteristics and the style of discussion do affect the accuracy. Addressing the semantic aspects of common terms will make the measures more reliable, and studies are ongoing to determine how to include these in our model in order to provide better document recommendations. Identifying additional document relationships and applying them in novel IR applications are interesting directions for future work.

References

- MIT (2012). Mit open courseware. <http://ocw.mit.edu/>.
- NPTEL (2012). National Programme on Technology Enhanced Learning, NPTEL. <http://nptel.iitm.ac.in/>.
- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches., *Proceedings of Human Language Technologies: NAACL*, Association for Computational Linguistics (pp. 19–27).
- Aletras, N., Stevenson, M., & Clough, P. (2012). Computing similarity between items in a digital library of cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(4), 16.
- Andrews, K., Gütl, C., Moser, J., Sabol, V., & Lackner, W. (2001). Search result visualisation with xfind., *User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on*, IEEE (pp. 50–58).
- Balogopalan, A., Balasubramanian, L.L., Balasubramanian, V., Chandrasekharan, N., & Damodar, A. (2012). Automatic keyphrase extraction and segmentation of video lectures., *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*, IEEE (pp. 1–10).
- Bean, A., & Green, R. (2001). *Relationships in the Organization of Knowledge* Vol. 2. Berlin: Springer.
- Capelle, M., Hogenboom, F., Hogenboom, A., & Frasincar, F. (2013). Semantic news recommendation using wordnet and bing similarities., *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ACM (pp. 296–302).
- Chalmers, M., & Chitson, P. (1992). Bead: Explorations in information visualization., *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, ACM (pp. 330–337).
- Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM*, 48(12), 152–152.
- Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2–3), 285–307.
- Frantzi, K.T., & Ananiadou, S. (1996). Extracting nested collocations., *Proceedings of the 16th conference on Computational linguistics-Volume 1* (pp. 41–46).

- Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115–130.
- Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJCAI*, (Vol. 7 pp. 1606–1611).
- Gonzalez-Agirre, A., Rigau, G., Agirre, E., Aletras, N., & Stevenson, M. (2015). Why are these similar? Investigating item similarity types in a large digital library. *Journal of the Association for Information Science and Technology*.
- Gouws, S. (2010). Evaluation and development of conceptual document similarity metrics with content-based recommender applications, Stellenbosch: University of Stellenbosch.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296): Morgan Kaufmann Publishers Inc.
- Hopfgartner, F. (2010). Personalised video retrieval: Application of implicit feedback and semantic user profiles, University of Glasgow.
- Huang, A. (2008). Similarity measures for text document clustering., *Proceedings of the Sixth New Zealand Computer Science Research Student Conference* (pp. 49–56).
- Huang, L., Milne, D., Frank, E., & Witten, I.H. (2012). Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology*, 63(8), 1593–1608.
- Huynh, T., Hoang, K., Do, L., Tran, H., Luong, H., & Gauch, S. (2012). Scientific publication recommendations based on collaborative citation networks., *Collaboration Technologies and Systems (CTS), 2012 International Conference on* (pp. 316–321).
- Khoo, C.S.G., & Na, J.C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, 40, 157–228.
- Lai, C.H., Liu, D.R., & Lin, C.S. (2013). Novel personal and group-based trust models in collaborative filtering for document recommendation. *Information Sciences*, 239(0), 31–49.
- McCormack, A.J., & Yager, R.E. (1989). A new taxonomy of science education. *Science Teacher*, 56(2), 47–48.
- Rafi, M., & Shaikh, M.S. (2013). An improved semantic similarity measure for document clustering based on topic maps. arXiv:1303.4087.
- Schaefer, C., Hienert, D., & Gottron, T. (2014). Normalized Relevance Distance—A Stable Metric for Computing Semantic Relatedness over Reference Corpora, ECAI.
- Strube, M., & Ponzetto, S.P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia., *AAAI*, (Vol. 6 pp. 1419–1424).
- Turdakov, D., & Velikhov, P. (2008). Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation.
- Wan, X. (2007). A novel document similarity measure based on earth mover's distance. *Information Sciences*, 177(18), 3718–3730.
- Wan, X.J., & Peng, Y.X. (2005). A new retrieval model based on texttiling for document similarity search. *Journal of Computer Science and Technology*, 20(4), 552–558.
- Wu, H.C., Luk, R.W.P., Wong, K.F., & Kwok, K.L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3), 13.
- Zarrinkalam, F., & Kahani M. (2012). *A new metric for measuring relatedness of scientific papers based on non-textual features*: Scientific Research Publishing.
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an efficient data clustering method for very large databases., *ACM SIGMOD Record*, (Vol. 25 pp. 103–114).