Improving Cross-Document Knowledge Discovery Through Content and Link Analysis of Wikipedia Knowledge

Peng Yan(✉) and Wei Jin

Department of Computer Science, North Dakota State University,
1340 Administration Ave., Fargo, ND 58102, USA
{peng.yan,wei.jin}@ndsu.edu

Abstract. The Vector Space Model (VSM) has been widely used in Natural Language Processing (NLP) for representing text documents as a Bag of Words (BOW). However, only document-level statistical information is recorded (e.g., document frequency, inverse document frequency) and word semantics cannot be captured. Improvement towards understanding the meaning of words in texts is a challenging task and sufficient background knowledge may need to be incorporated to provide a better semantic representation of texts. In this paper, we present a text mining model that can automatically discover semantic relationships between concepts across multiple documents (where the traditional search paradigm such as search engines cannot help much) and effectively integrate various evidences mined from Wikipedia knowledge. We propose this integration may effectively complement existing information contained in text corpus and facilitate the construction of a more comprehensive representation and retrieval framework. The experimental results demonstrate the search performance has been significantly enhanced against two competitive baselines.

Keywords: Knowledge discovery · Semantic relatedness · Cross-Document knowledge discovery · Document representation

1 Introduction

Text mining aims at mining high-quality information from mass text. However, great challenges have been posed for many text mining tasks because of the increasing sheer volume of text data and the difficulty of capturing valuable knowledge hidden in them. Therefore efficient and high-quality text mining algorithms are demanded and effective document representation and accurate semantic relatedness estimation become increasingly crucial. Traditional approaches for document representation are mostly based on the Vector Space (VSM) Model or the Bag of Words (BOW) model which takes a document as an unordered collection of words and only document-level

This submission is an extended version of the paper published in DaWaK'12, which was selected by the DaWaK'12 program committee for possible publication in the LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems.

A. Hameurlain et al. (Eds.): TLDKS XXI, LNCS 9260, pp. 161–184, 2015.
DOI: 10.1007/978-3-662-47804-2_8

statistical information is recorded (e.g., document frequency, inverse document frequency). Due to the lack of capturing semantics in texts, for certain tasks, especially fine-grained information discovery applications, such as mining relationships between concepts, VSM demonstrates its inherent limitations because of its rationale for computing relatedness between words only based on the statistical information collected from documents themselves. It leads to great semantic loss because terms not appearing in the text literally cannot be taken into consideration.

Our previous work [1] introduced a special case of text mining focusing on detecting semantic relationships between two concepts across documents, which we refer to as Concept Chain Queries (CCQ). A concept chain query involving concepts A and B has the following meaning: find the most plausible relationship between concept A and concept B assuming that one or more instances of both concepts occur in the corpus, but not necessarily in the same document. For example, both may be football lovers, but mentioned in different documents. However, the previous solution was built under the VSM assumption only for the document collection, which limited the scope of the discovered results. For instance, "Albert Gore" is closely related to "George W. Bush" since the two men together produced the most controversial presidential election in 2000, which was the only time in American history that the Supreme Court has determined the outcome of a presidential election. However, "Albert Gore" cannot be identified as a relevant concept to "George W. Bush" if it does not occur in the document collection where the concept chain queries are performed. Furthermore, the semantic relatedness between concepts computed in [1] is solely measured by the statistical information gathered from the corpus such as term frequency (TF), inverse document frequency (IDF), with no background knowledge incorporated.

In this work, we present a new approach that attempts to address the above problems by utilizing background knowledge to provide a better semantic representation of any text and a more appropriate estimation of semantic relatedness between concepts. This is accomplished through leveraging Wikipedia, the world's currently largest human built encyclopedia. Specifically, in addition to inspecting the given documents, we sift through the articles and anchor texts in the space of Wikipedia, attempting to integrate relevant background knowledge for the topics being searched. Our algorithm is motivated by the Explicit Semantic Analysis (ESA) [3] technique where ESA maps a given text or concept to a conceptual vector space spanned by all Wikipedia articles, and thus rich background knowledge can be integrated into the semantic representation of that text or concept. Here we adapt and improve the ESA model in two dimensions. First, we attempt to identify only the most relevant concepts generated from ESA for topic semantic representation and relatedness computation through introducing a series of heuristic steps for noise reduction. Second, we go one step further to take into account anchor texts inside relevant Wikipedia articles. This is based on the observation that the anchor texts within an article are usually highly relevant to it. Therefore, if an article is identified to be relevant to our search topic, it is highly likely that its anchors are topic-relevant as well. To validate the proposed techniques, a significant amount of queries covering different scenarios were conducted. The results showed that through incorporating Wikipedia knowledge, the most relevant concepts to the given topics were ranked in top positions.

Our contribution of this effort can be summarized as follows: (1) a new Wiki-enabled cross-document knowledge discovery framework has been proposed and implemented which effectively complements the existing information contained in the document collection and provides a more comprehensive knowledge representation and mining framework supporting various query scenarios; (2) effective noise filtering techniques are provided through developing a series of heuristic strategies for noise reduction, which further increases the reliability of the overall knowledge encoded; (3) to the best of our knowledge, little work has been done to consider ESA as an effective aid in cross-document knowledge discovery. In this work, built on the traditional BOW representation for corpus content analysis, the ESA technique has been successfully integrated into the discovery process and a better estimation of semantic relatedness is provided by combining various evidences from Wikipedia such as article content and anchor texts. We envision this integration would also benefit other related tasks such as question answering and cross–document summarization; (4) the proposed approach presents a new perspective of alleviating semantic loss caused by only using the Vector Space Model (VSM) on the corpus level through incorporating relevant background knowledge from Wikipedia; (5) in addition to uncovering "what relationships might exist between two topics of interest", our method further explores another dimension of the analysis by generating evidence trails from Wikipedia to interpret the nature of the potential concept relationships.

The remainder of this paper is organized as follows: Sect. 2 describes related work. Section 3 introduces concept chain queries. In Sect. 4, we present our proposed method of utilizing Wikipedia knowledge for answering concept chain queries. Experimental results are presented and analysed in Sect. 5, and is followed by the conclusion and future work.

2 Related Work

Mining semantic relationships/associations between concepts from text is important for inferring new knowledge and detecting new trends. Built within the discovery framework established by Swanson and Smalheiser [4], Srinivasan proposed the open and closed text mining algorithm [2] to automatically discover interesting concepts from MEDLINE. There has also been work on discovering connections between concepts across documents using social network graphs, where nodes represent documents and links represent connections (typically URL links) between documents. However, much of the work on social network analysis has focused on special problems, such as detecting communities [7, 11]. Our previous work [1] introduced Concept Chain Queries (CCQ), a special case of text mining focusing on detecting cross-document links between concepts in document collections, which was motivated by Srinivsan's closed text mining algorithm [4]. Specifically, the solution proposed attempted to generate concept chains based on the "Bag of Words" (BOW) representation on the text corpus and extended the technique in [2] by considering multiple levels of interesting concepts instead of just one level as in the original method. Each document in [1] was represented as a vector containing all the words appearing in the relevant text snippets in the corpus but did not take any auxiliary knowledge into

consideration, whereas in this new solution, in addition to corpus level content analysis, we further examine the potential of integrating the Explicit Semantic Analysis (ESA) [3] technique to better serve this task which effectively incorporates more comprehensive knowledge from Wikipedia. There have been a lot of efforts in earlier research as discussed in [31], trying to add semantics to traditional VSM based text processing. Deerwester [32] introduced Latent Semantic Indexing (LSI) for automatic identification of concepts using singular value decomposition. However, it has been found that LSI can rarely improve the strong baseline established by SVM [5, 35, 36]. This becomes part of our motivations of integrating ESA in this work.

WordNet, a lexical database for the English language [18], has been widely used to overcome the limitations of the VSM in text retrieval [19], document clustering [20, 21] and document categorization [22, 23]. For example, Hotho et al. [6] utilized WordNet to improve the VSM text representation and Scott et al. [9] proposed a new representation of text based on WordNet hypernyms. These WordNet-based approaches were shown to alleviate the problems of BOW model but are subject to relatively limited coverage compared to Wikipedia, the world's largest knowledge base to date. Gurevych et al. used Wikipedia to integrate semantic relatedness into the information retrieval process [24], and Müller et al. [25] used Wikipedia in domain-specific information retrieval. Gabrilovich et al. [5] applied machine learning techniques to Wikipedia and proposed a new method to enrich document representation from this huge knowledge repository. Specifically, they built a feature generator to identify most relevant Wikipedia articles for each document, and then used concepts corresponding to these articles to create new features. As claimed in [5], one of the advantages using Wikipedia over Open Directory Project (ODP) is the articles in Wikipedia are much cleaner than typical Web pages, and mostly qualify as standard written English. However, without proper feature selection strategies employed, there will still be a large amount of noise concepts introduced by the feature generator. Another concern needing to be drawn here is the challenge of efficiently processing large scale data. Bonifati and Cuzzocrea [28] presented a novel technique to fragment large XML documents using structural constraints such as size, tree-width, and tree-depth. Cuzzocrea et al. [29] used K-means clustering algorithm to perform the fragmentation of very large XML data warehouses at scale. Cuzzocrea and Bertino [30] proposed a framework for efficiently processing distributed collections of XML documents. While in this work, we import the XML dump of Wikipedia into relational database and build multi-level indices on the database to support efficient queries against Wiki data.

Serving as an integral part of information retrieval and natural language processing, semantic similarity estimation between words has gained increasing attention over the past years. Various web resources have been considered for this purpose [14–16]. Rinaldi [34] proposed a metric to compute the semantic relatedness between words based on a semantic network built from ontological information. Bollegala [17] developed an automatic method for semantic similarity calculation using returned page counts and text snippets generated by a Web search engine. Gabrilovich et al. also [3] presented a novel method, Explicit Semantic Analysis (ESA), for fine-grained semantic representation of unrestricted natural language texts. Using this approach, the meaning of any text can be represented as a weighted vector of Wikipedia-based concepts (articles), called an interpretation vector [3]. Gabrilovich et al. [3] also discussed the

problem of possibly containing noise concepts in the vector, especially for text fragments containing multi-word phrases (e.g., multi-word names like George Bush). Our proposed solution is motivated by this work and to tackle the above problems we have developed a sequence of heuristic strategies to filter out irrelevant concepts and clean the vector. Another interesting work is an application of ESA in a cross-lingual information retrieval setting to allow retrieval across languages [8]. In that effort the authors performed article selection to filter out those irrelevant Wikipedia articles (concepts). However, we observed the selection process resulted in the loss of many dimensions in the following mapping process, whereas in our proposed approach, the process of article selection is postponed until two semantic profiles have been merged so that the semantic loss could be possibly reduced to the minimum. Furthermore, in comparison to [13, 27], we also tap into another valuable information resource, i.e. the Wikipedia anchor texts, along with articles to provide better semantic relatedness estimation.

3 Concept Chain Queries

As described earlier, concept chain query (CCQ) is attempting to detect links between two concepts (e.g., two person names) across documents. A concept chain query involving concept A and concept B intends to find the best path linking concept A to concept B. The paths found stand for potential conceptual connections between them. Figure 1 gives an example of CCQ, where the query pair is "Nashiri :: Nairobi attack". Since "Nashiri" co-occurs with "Jihad Mohammad Ali al Makki" in the same sentence in Document 1, and "Nairobi attack" co-occurs with "Jihad Mohammad Ali al Makki" in the same sentence in Document 2, "Nashiri" and "Nairobi attack" can be linked through the concept "Jihad Mohammad Ali al Makki".

Document 1:
Nashiri and *his cousin*, *Jihad Mohammad*, returned to Afghanistan, probably in 1997, Nashiri again encountered Bin Ladin, still recruiting for "the coming battle with the United States." Nashiri joined al Qaeda and later was recognized as the chief of al Qaeda operations in and around the Arabian Peninsula.
Document 2:
In late 1998, al Qaeda decided mounting an attack against a U.S. vessel and *Jihad Mohammad*, also known as *Azzam*, was a suicide bomber for the ***Nairobi attack***.

Fig. 1. A concept chain example for the query "Nashiri :: Nairobi attack"

3.1 Semantic Profile for Topic Representation

A semantic profile is essentially a set of concepts that together represent the corresponding topic. To further differentiate between the concepts, semantic type (ontological information) is employed in profile generation. The concept mapping process is basically a two-step task: (1) we extract concepts from the document collection using

Semantex [10]; (2) the extracted concepts are mapped the counterterrorism domain ontology [1]. Table 1 illustrates part of semantic type – concept mappings.

Table 1. Semantic type - concept mapping

Semantic type	Instances
Religion	Islam, Muslim
Human action	attack, killing, covert action, international terrorism
Leader	vice president, chief, governor
Country	Iraq, Afghanistan, Pakistan, Kuwait
Infrastructure	World Trade Centre
Diplomatic building	consulate, pentagon, UAE Embassy

Thus each profile is defined as a vector composed of a number of semantic types.

$$profile(T) = \{ST_1, ST_2, \ldots, ST_n\} \tag{1}$$

Where ST_i represents a semantic type to which concepts appearing in the topic-related text snippets belong. We used sentence as window size to measure relevance of appearing concepts to the topic term. Under this representation each semantic type is again referred to as an additional level of vector composed of a number of terms that belong to this semantic type.

$$ST_i = \{w_{i,1}m_1, w_{i,2}m_2, \ldots, w_{i,n}m_n\} \tag{2}$$

Where m_j represents a concept belonging to semantic type ST_i, and $w_{i,j}$ represents its weight under the context of ST_i and sentence level closeness. When generating the profile we replace each semantic type in (1) with (2).

In (2), to compute the weight of each concept, we employ a variation of *TF*IDF* weighting scheme and then normalize the weights:

$$w_{i,j} = s_{i,j} \,/\, highest(s_{i,l}) \tag{3}$$

Where $l = 1, 2, \ldots, r$ and there are totally r concepts for ST_i, $s_{i,j} = df_{i,j}{*}Log(N/df_j)$, where N is the number of sentences in the collection, df_j is the number of sentences concept m_j occurs, and $df_{i,j}$ is the number of sentences in which topic T and concept m_j co-occur and m_j belongs to semantic type ST_i. By using the above three formulae we can build the corresponding profile representing any given topic.

3.2 Concept Chain Generation

We adapt Srinivasan's closed discovery algorithm [2] to build concept chains for any two given topics. Each concept chain generated reveals a plausible path from concept A to concept C (suppose A and C are two given topics of interest). The algorithm of generating concept chains connecting A to C is composed of the following three steps.

1. Conduct independent searches for A and C. Build the A and C profiles. Call these profiles AP and CP respectively.
2. Compute a B profile (BP) composed of terms in common between AP and CP. The weight of a concept in BP is the sum of its weights in AP and CP. This is the first level of intermediate potential concepts.
3. Expand the concept chain using the created BP profile together with the topics to build additional levels of intermediate concept lists which (i) connect the topics to each concept in BP profile in the sentence level within each semantic type, and (ii) also normalize and rank them (as detailed in Sect. 3.1).

4 Wikipedia as an Information Resource

Wikipedia is currently the largest human built encyclopedia in the world. It has over 5,000,000 articles by April 05, 2011, and is maintained by over 100,000 contributors from all over the world. As of February 2013, there are editions of Wikipedia in 285 languages. Knowledge in Wikipedia ranges from psychology, math, physics to social science and humanities. To utilize Wikipedia knowledge to complement the existing information contained in the document collection, two important information resources, Wikipedia article contents and anchor texts are considered. Specifically, appropriate content and link analysis will be performed on Wikipedia data and the mined relevant knowledge will be used to further improve our query model and semantic relatedness estimation module.

4.1 Semantic Relatedness Measures

Semantic relatedness indicates degree to which words are associated via any type (such as synonymy, meronymy, hyponymy, hypernymy, functional, associative and other types) of semantic relationships [37]. The measures of computing semantic relatedness between concepts can be grouped into four classes in general [33]: the path length based measures that use the length of path connecting concepts in the taxonomy to measure the closeness between concepts; the information content based measures that rely on the shared information content between concepts; the feature based measures that exploit the common characteristics of concepts; and the hybrid measures that combine the previous three measures. The similarity measures defined in this work can be viewed as an extension of the information content based measure.

4.2 Article Content Analysis

For content analysis, we have adapted the Explicit Semantic Analysis (ESA) technique proposed by Gabrilovich et al. [3] as our underlying content-based measure for analyzing Wikipedia articles relevant to the given topics of interest. In ESA, each term (e.g., topic of interest) is represented by a concept vector containing relevant concepts (Wikepedia articles) to the topic along with their association strengths and each text

fragment can also be mapped to a weighted vector of Wikipedia concepts called an interpretation vector. Therefore, computing semantic relatedness between any two text fragments can be naturally transformed into computing the Cosine similarity between interpretation vectors of two texts.

Using the ESA method, each article in Wikipedia is treated as a Wikipedia concept (the title of an article is used as a representative concept to represent the article content), and each document is represented by an interpretation vector containing related Wikipedia concepts (articles) with regard to this document. Formally, a document d can be represented as follows:

$$\phi(d) = < as(d, a_1), \ldots, as(d, a_n) > \tag{4}$$

Where $as(d,a_i)$ denotes the association strength between document d and Wikipedia article a_i. Suppose d is spanned by all words appearing in it, i.e., $d =< w_1, w_2, \ldots, w_j >$, and the association strength $as(d,a_i)$ is computed by the following function:

$$as(d, a_i) = \sum_{w_j \in d} tf_d(w_j) tf \bullet idf_{a_i}(w_j) \tag{5}$$

Where $tf_d(w_j)$ is the occurrence frequency of word w_j in document d, and $tf \bullet idf_{a_i}(w_j)$ is the tf•idf value of word w_j in Wikipedia article a_i. As a result, the vector for a document is represented by a list of real values indicating the association strength of a given document with respect to Wikipedia articles. By using efficient indexing strategies such as single-pass in memory indexing, the computational cost of building these vectors can be reduced to within 200–300 ms. In concept chain queries, the topic input is always a single concept (a single term or phrase), and thus Eq. (5) can be simplified as below as $tf_d(w_j)$ always equals 1:

$$as(d, a_i) = \sum_{w_j \in d} tf \bullet idf_{a_i}(w_j) \tag{6}$$

As discussed above, the original ESA method is subject to the noise concepts introduced, especially when dealing with multi-word phases. For example, when the input is *Angelina Jolie*, the generated interpretation vector will contain a fair amount of noise concepts such as *Eudocia Angelina*, who was the queen consort of Stephen II Nemanjić of Serbia from 1196 to 1198. This Wikipedia concept (article) is selected and ranked high in the interpretation vector because the term Angelina occurs many times in the article "*Eudocia Angelina*", but obviously this article is irrelevant to the given topic *Angelina Jolie*.

In order to make the interpretation vector more precise and relevant to the topic, we have developed a sequence of heuristics to clean the vector. Basically, we use a modified Levenshtein Distance algorithm to measure the relevance of the given topic to each Wikipedia concept generated in the interpretation vector. Instead of using allowable edit operations of a single character to measure the similarity between two strings as in the original Levenshtein Distance algorithm, we view a single word as a unit for edit operations, and thus the adapted algorithm can be used to compute the

similarity between any two text snippets. The heuristic steps used to remove noise concepts are illustrated in Fig. 2.

Input: a topic T of interest
an interpretation vector V representing the topic T

Output: a cleaned Wikipedia-based concept vector V' representing the topic T

1. If T is a single word topic, then count the number of occurrences of T in the article texts represented by each concept v_i in V, respectively. If T occurs more than 3 times, then keep v_i in V, otherwise, remove v_i from V.

2. If T is a multi-word topic, then the adapted Levenshtein distance algorithm applies to measure the relevance of each Wikipedia concept (article) v_i in V to topic T.

2.1. If NumOfWords(T) $\leq$ 2, then extract all text snippets TS_j within the window size NumOfWords(T)+1 from the article text of v_i. If there exists a j such that LevenshteinDistance(T, TS_j) <= 1, then keep v_i in V, otherwise, remove v_i from V.

2.2. If NumOfWords(T) > 2, then extract all text snippets TS_j within the window size NumOfWords(T)+2 from the article text of v_i. If there exists j such that LevenshteinDistance(T, TS_j) $\leq$ 2, then keep v_i in V, otherwise, remove v_i from V.

Fig. 2. The interpretation vector cleaning procedure

4.3 Article Link Analysis

Anchor texts, another type of valuable information resource provided by Wikipedia in addition to the textual content of articles, imply rich hidden associations between different Wikipedia concepts. For example, the Wikipedia article talking about "*Osama bin Laden*" contains a great number of potential terrorists who are related to him and terrorism events that he was involved in (appearing as anchor texts in the article). Therefore, through inspecting anchor texts in each relevant Wikipedia article, we are able to find a fair amount of interesting concepts related to the topic. Figure 3 gives part of the anchors in the article "*Osama bin Laden*".

We assume that two concepts (articles) sharing similar anchors may be closer to each other in terms of semantic relatedness. As discussed earlier, given a topic of interest, we can represent it as an interpretation vector containing the relevant Wikipedia articles using the ESA method. Also, each Wikipedia article can be further represented by the anchors appearring in it. Therefore, we can build an additional vector, called anchor vector, based on the interpretation vector produced for a given search topic. Simiarly, we can approach the semantic relatedness between two topics from another perspective by calculating the Cosine score of the two anchor vectors built for them.

Formally, suppose the interpretation vector for a topic T_i is $V_i =< article_1, article_2, \ldots, article_m >$, where $article_i$ in V_i represents a Wikipedia article relevant to T_i, then the topic T_i can be further represented as an *Anchor Vector (AV)* as follows.

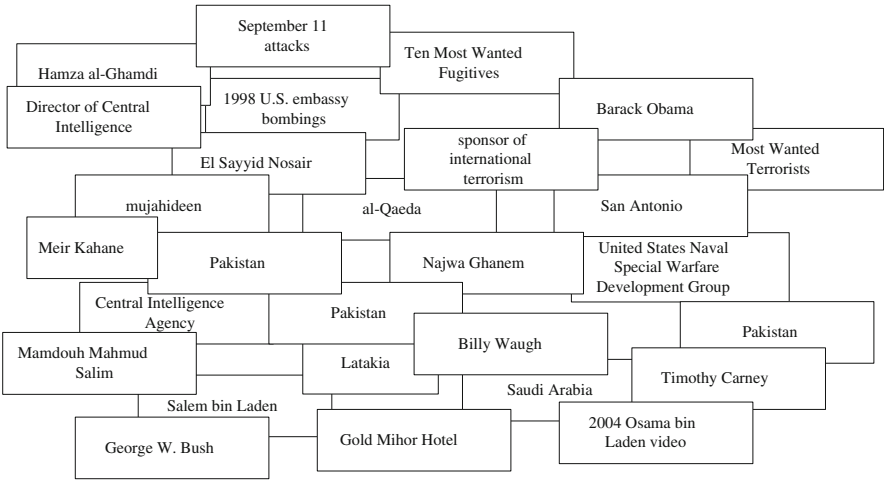

Fig. 3. Wikipedia anchors related to "*Osama bin Laden*"

$$\begin{aligned} AV(T_i) = << & w_{i,1,1}anchor_{1,1}, w_{i,2,1}anchor_{2,1}, \ldots >, \\ & \ldots, \\ & < w_{i,1,m}anchor_{1,m}, w_{i,2,m}anchor_{2,m}, \ldots > > \end{aligned} \tag{7}$$

Where $anchor_{x,y}$ represents the anchor text $anchor_x$ appearing in $article_y$ in V_i, and $w_{i,x,y}$ is the weight for $anchor_{x,y}$. To calculate $w_{i,x,y}$, we count the number of sub-vectors within $AV(T_i)$ in which $anchor_{x,y}$ appears, and then normalize it:

$$w_{i,x,y} = \frac{w_{i,x,y}}{highest(w_{i,d,y})} \tag{8}$$

Where d = 1,2,…,r and there are totally r anchors in Wikipedia. Therefore, the semantic relatedness between two topics of interest can be estimated as follows:

$$Sim(T_i, T_j) = Cosine(AV(T_i), AV(T_j)) \tag{9}$$

4.4 Integrating Wikipedia Knowledge into Concept Chain Queries

Given the advantages of using Wikipedia as an effective information aid for semantic representation, we integrate the knowledge derived from Wikipedia into our concept chain queries. Specifically, we build interpretation vectors (using our adapted ESA method) and anchor vectors (using the method described in Sect. 4.2) for both the two given topics and each intermediate concept in the merged BP profile, and then compute the Cosine similarities between the topics and each concept in the BP profile using the corresponding interpretation vectors and anchor vectors, respectively. The final ranking will be an integrated scheme considering the following three types of similarities.

Corpus-level TF*IDF-based Similarity. As the most widely used document representation, the BOW representation has demonstrated its advantages. It is simple to compute and strictly sticking to the terms occurring in the document, thereby preventing outside noise concepts that do not appear in the document from flowing into the feature space of the representation. Given these benefits, a variation of *TF*IDF* weighting scheme under the context of BOW representation is incorporated into our final ranking to capture corpus level statistical information. We call this kind of similarity the TF*IDF-based similarity.

ESA-based Similarity. Unlike the BOW model, ESA makes use of the knowledge outside the documents themselves to compute semantic relatedness. It well compensates for the semantic loss resulted from the BOW technique. The relatedness between two concepts in ESA is computed using their corresponding interpretation vectors containing related concepts derived from Wikipedia. In the context of concept chain queries, we compute the Cosine similarity between the interpretation vectors of topic A and each concept in the intermediate BP profile, as well as between topic C and each concept V_i, and take the average of two Cosine similarities as the overall similarity for each concept V_i in BP. We call this kind of similarity the ESA-based similarity.

Anchor-based Similarity. Anchor texts have served as another important information aid in our algorithms to provide highly relevant concepts to the given topics through considering the descriptive or contextual information for relevant Wikipedia articles. As with the case of computing the ESA-based similarity for topic A(C) and each concept V_i in the intermediate BP profile using the interpretation vectors, here anchor vectors are used to measure concept closeness. We refer to this type of similarity the Anchor-based similarity.

Integrating TF*IDF-based Similarity, ESA-based Similarity and Anchor-based Similarity into the Final Ranking. The TF*IDF-based similarity, ESA-based similarity and Anchor-based similarity are finally combined to form a final ranking for concepts generated in the intermediate profiles:

$$S_{overall} = (1 - \lambda_1 - \lambda_2)S_{TFIDF} + \lambda_1 S_{ESA} + \lambda_2 S_{anchor} \tag{10}$$

Where λ_1 and λ_2 are two tuning parameters that can be adjusted based on the preference on the three similarity schemes in the experiments. S_{TFIDF} refers to the TF*IDF-based similarity, S_{ESA} the ESA-based similarity, and S_{anchor} the Anchor-based similarity.

4.5 Annotating Semantic Relationships Between Concepts

In addition to answering "what relationships might exist between two topics?", we go one step further to collect relevant text snippets extracted from multiple Wikipedia articles in which the discovered chains appear. This is in fact a multi-document summary that explains the plausible relationship between topics with intensive knowledge derived from Wikipedia. For example, given a query pair: "*Bin Laden*" and "*Abdel-Rahman*", one of the discovered concept chains is: *Bin Laden* → *Azzam* → *Abdel-Rahman*. Our goal now is to find supporting evidence that interprets how

"*Bin Laden*" is linked to "*Abdel-Rahman*" through "*Azzam*" in the space of Wikipedia. We consider this process as the chain-focused sentence retrieval problem and decompose it into the following two subtasks.

Chain-Relevant Article Retrieval. This subtask takes a generated concept chain as input and attempts to find relevant Wikipedia articles for it. One important criterion that needs to be met is the identified Wikepeida articles should be relevant to the whole chain (i.e. relevance to the given topics (end points of the chain) as well as intervening concepts), not just to any individual segment of the chain. To achieve this, we first (1) build the corresponding interpretation vectors for all of the concepts appearing in the chain; (2) perform noise removal using the cleaning procedure described in Fig. 2; (3) construct a ranked list of Wikipedia articles by intersecting the resulting interpretation vectors with each article weighted using formula 6; (4) follow similar steps as above to construct a ranked list of anchors (note that an anchor also represents a Wikipedia article) with each anchor weighted using formula 8. The articles represented by the concepts in the two ranked lists are viewed as chain relevant articles.

Chain-Focused Sentence Retrieval. This step inspects the content of each article generated from the previous step and extracts sentences that explain each segment of chain. For example, the chain *Bin Laden* → *Azzam* → *Abdel-Rahman* is composed of two segments: *Bin Laden* → *Azzam* and *Azzam* → *Abdel-Rahman*. For the segment *Bin Laden* → *Azzam*, sentences where "*Bin Laden*" and "*Azzam*" co-occur will be extracted as supporting evidence for this partial chain. Figure 4 shows the generated evidence trail for this example.

Supporting Evidence for ***Bin Laden*** and ***Azzam***
In1989,after the Soviets pulled out of Afghanistan, ***Azzam*** *and his deputy Osama* ***bin Laden*** *decided to keep their movement permanent and founded the Al Qaeda.*

Supporting Evidence for ***Azzam*** and ***Abdel-Rahman***
During theological studies in Egypt, ***Azzam*** *met Omar* ***Abdel-Rahman****,Dr.Ayman al-Zawahiri and other followers of Sayyed Qutb,an extremely influential leader of the Egyptian Muslim Brotherhood,who had been executed by President Gamal Abdel Nasser in 1966.*

Fig. 4. Evidence trail generated from Wikipedia articles for the concept chain *Bin Laden* → *Azzam* → *Abdel-Rahman*

4.6 The New Mining Model

To summarize, the new model of answering concept chain queries consists of two sequential steps as shown in Figs. 5 and 6. Figure 5 illustrates the first step which discovers potential relationships between two given topics from the given document collection without background knowledge incorporated. Figure 6 details how Wikipedia knowledge is integrated into this discovery process and facilitates better estimation of semantic relatedness between concepts. Also, we go one step further and require the response to be a set of Wikipedia text snippets (i.e. evidence trail) in which

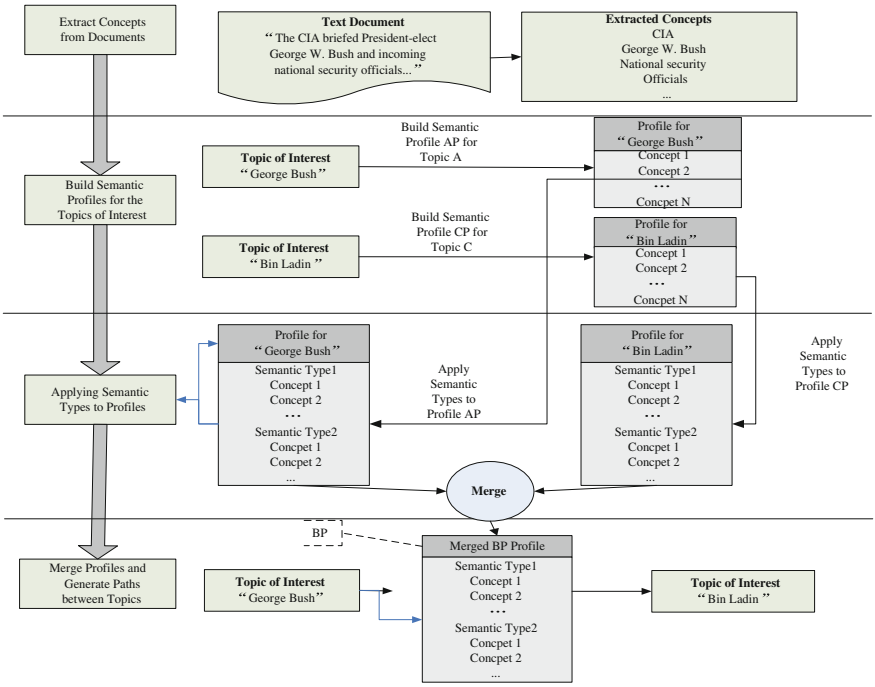

Fig. 5. The new model of answering concept chain queries: component-1

the discovered concept chain occurs. This may assist a user with the second dimension of the analysis process, i.e. when the user has to peruse the documents to figure out the nature of the relationship underlying a suggested chain.

5 Empirical Evaluation

A challenging task for the evaluation was constructing an evaluation data set, since there are no standard data sets available for quantitatively evaluating concept chains. We performed our evaluation using the 9/11 counterterrorism corpus. The Wikipedia snapshot used in the experiments was dumped on April 05, 2011.

5.1 Processing Wikipedia Dumps

As an open source project, the entire content of Wikipedia is easily obtainable. All the information from Wikipedia is available in the form of database dumps that are released periodically, from several days to several weeks apart. The version used in this work was released on April 05, 2011, which was separated into 15 compressed XML files and totally occupies 29.5 GB after decompression, containing articles, templates, image descriptions, and primary meta-pages. We leveraged MWDumper [12] to import the XML dumps into our MediaWiki database, and after the parsing process, we identified 5,553,542 articles.

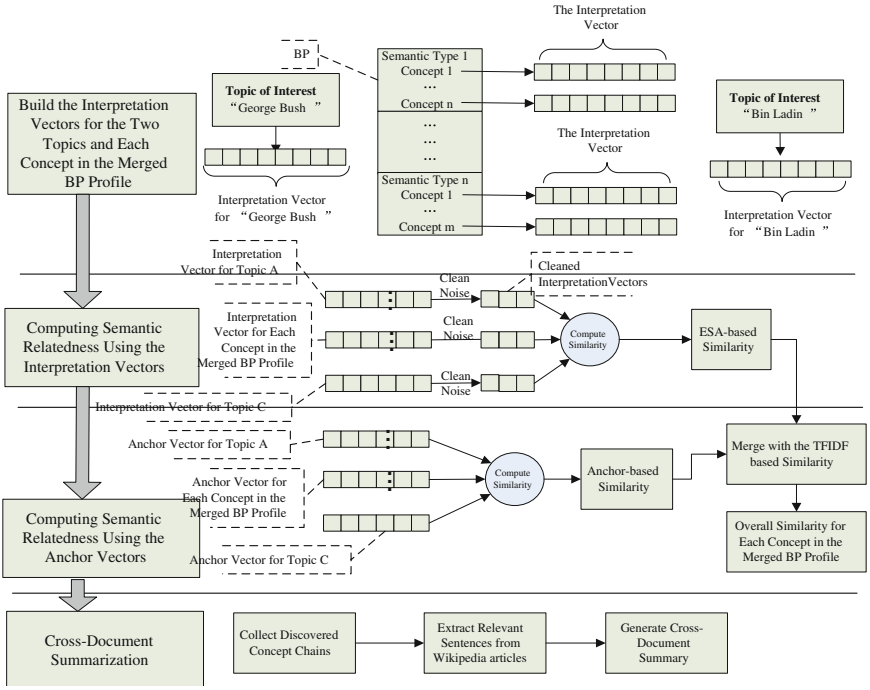

Fig. 6. The new model of answering concept chain queries: component-2

5.2 Evaluation Data

We performed concept chain queries on the 9/11 counterterrorism corpus. This involves processing a large open source document collection pertaining to the 9/11 attack, including the publicly available 9/11 commission report. The report consists of Executive Summary, Preface, 13 chapters, Appendix and Notes. Each of them was considered as a separate document resulting in 337 documents. The whole collection was processed using Semantex [10] and concepts were extracted and selected as shown in Table 1. Query pairs covering various scenarios (e.g., ranging from popular entities to rare entities) were selected by the assessors and used as our evaluation data. We selected chains of lengths ranging from 1 to 4 in terms of the number of associations. The chains were selected by going through the same procedure as in [26], which is also described as follows:

1. We ran queries with various pairs of topics: in the counterterrorism corpus, the topics were mostly named entities.
2. For each topic pair, the relevant paragraphs for either topic were then manually inspected: we selected those where there was a logical connection between the two topics.
3. After achieving agreement among all annotators, we then generated the concept chains for these topic pairs (and paragraphs) as evaluation data.

The above process generated 37 chains in 9/11 corpus which will be used as truth chains for later experiments.

5.3 Experimental Results

Parameter Settings. As mentioned in Sect. 4.3, a combination of TF*IDF-based similarity, ESA-based similarity and Anchor-based similarity is used to rank the links detected by our system. λ_1 and λ_2 in Eq. 10 are two parameters that need to be tuned so that the generated similarity between two concepts best matches the judgements from our assessors. To accomplish this, we first built a set of training data composed of 10 query pairs randomly selected from the evaluation set, and then generated BP profiles for each of them using our proposed method. Among each BP profile, we selected the top 5 concepts (links) within each semantic type, and compared their rankings with the assessors' judgements. The values of λ_1 and λ_2 were tuned in the range of [0.1, 1]. Specifically, we set $\lambda_2 = 0$ or $\lambda_1 = 0$ to evaluate the contribution of each individual part (the ESA-based similarity or the Anchor-based similarity) in the final weighting scheme. When $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, the best performance was obtained when $\lambda_1 = 0.4$ and $\lambda_2 = 0.3$. These settings were also used in our later experiments.

Query Results. Before proceeding to the evaluation of the proposed model, we first conducted an experiment to demonstrate the improved performance of our adapted ESA method against the original ESA. We selected 10 concepts that we have good knowledge about as shown in Table 2 and then built the interpretation vectors for each of them using the original ESA and our adapted ESA respectively. We calculated the averaged precision defined as below to measure the performance of the two approaches.

$$aveP = \left(\sum_{i=1}^{N} \frac{concept\ found\ and\ relevant}{total\ concepts\ found} \right) / N \tag{11}$$

Table 2. Ten concepts used for the interpretation vector construction

Semantic type	Belonging concept
Person	George Bush
	Bill Clinton
Organization	Central Intelligence Agency
	United States Federal Government
Event	World War
	September 11 attacks
	Lewinsky Scandal
Science	Data Mining
	Natural Language Processing
	Artificial Intelligence

where N is the number of concepts under consideration. The results are illustrated in Fig. 7 where the X-axis indicates the number of concepts kept in each of the interpretation vectors and the Y-axis indicates the averaged precision ratio. It is obvious that our adapted ESA achieves significant improvement over the original one for identifying topic-related Wikipedia concepts.

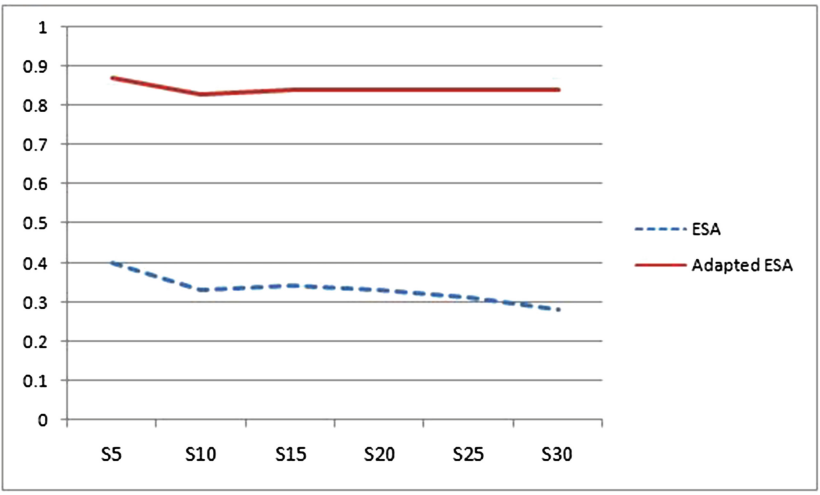

Fig. 7. The Averaged Precision of the generated interpretation vectors using the original ESA and adapted ESA based on processing data in Table 2

Table 3 shows the top 15 concepts generated in the interpretation vectors for 4 sample concepts. For example, for "Lewinsky Scandal", the top 15 concepts in the interpretation vector built using our adapted ESA include most of the people involved in this event in addition to Clinton and Lewinsky themselves, such as Linda Tripp who secretly recorded Lewinsky's confidential phone calls about her relationship with Clinton, and Betty Currie who was the personal secretary of Clinton and well known in the scandal for handling gifts given to Lewinsky by Clinton. However, most of the top concepts identified using the original ESA are representing some irrelevant events.

To further evaluate the performance of the original ESA and the adapted ESA in semantic profile generation, we selected 10 query pairs as shown in Table 4 and generated semantic profiles serving as linking concepts (i.e. BP profile) through selecting common concepts appearing in the two interpretation vectors built for the two given topics. Each concept in the semantic profile was weighted using the original ESA and our adapted ESA respectively. We again calculated the averaged precision to measure the percentage of the relevant concepts in the generated profile. The results are shown in Fig. 8 where the X-axis indicates the number of concepts kept in each generated semantic profile and the Y-axis indicates the averaged precision. It is demonstrated that for BP level semantic profile generation, our adapted ESA also performs much better than the original ESA.

Table 3. Top 15 concepts in the sample interpretation vectors using the adapted ESA and the original ESA

Input	#	Original ESA	Adapted ESA
Data Mining	1	Open-cast_mining	Relational_classification
	2	Opencast_Mining	Relational_data_mining
	3	Mining_engineer	Data_Mining_Extensions
	4	Open_cast_mining	Biological_data
	5	data	Java_Data_Mining
	6	Mine_(industry)	Weather_Data_Mining
	7	Open-cast_mine	National_Center_for_Data_Mining
	8	Golden_Source_of_data	Privacy_preserving_data_mining
	9	Data_withholding	Structure_mining
	10	Data_Havens	Oracle_Data_Mining
	11	Data_Warehousing	Cross_Industry_Standard_Process_for_Data_Mining
	12	Data_Transfer	Knowledge_discovery
	13	Data_rate_(disambiguation)	Data_Pre-processing
	14	Data_General_One	Data_mining_agent
	15	Data_matrix_(disambiguation)	Sequence_mining
Central Intelligence Agency	1	Agency_(disambiguation)	United_States._Central_Intelligence_ Agency
	2	United_States. _Central_Intelligence_Agency	Central_Intelligence_Agency_Museum
	3	Starfleet_Intelligence	Central_Intelligence_Agency_library
	4	Nigerian_intelligence	The_Agency
	5	Virginia_farmboys	National_Intelligence_Agency_(United_States)
	6	Directorate_for_Inter-Service_Intelligence	Agency
	7	Process_of_intelligence	Office_of_Scientific_Intelligence
	8	14th_Intelligence_Company	Intelligence_officer
	9	Intelligence_augmentation	Security_agency
	10	Human_intelligence_ (disambiguation)	John_N._McMahon
	11	Israeli_Intelligence_Agency	National_Intelligence_Board
	12	Agência_Brasileira_de_ Inteligência	Director_of_the_Central_Intelligence_Agency
	13	Central_(disambiguation)	Military_Intelligence_Division
	14	Administrative_agency	Private_intelligence_agency
	15	Job_agency	Intelligence_agency
Lewinsky Scandal	1	Scandal-mongering	Clinton:_His_Struggle_with_Dirt
	2	HIV-tainted-blood_scandal	Monica_Lewinsky
	3	Scandal_of_Scientology	Lewinsky_scandal
	4	The_Scandal_of_Scientology_ (book)	Linda_Tripp
	5	Iraq_War_Scandal_ (disambiguation)	Susan_Schmidt
	6	CDU_contribution_scandal	Kramerbooks_&_Afterwords
	7	Parmalat_scandal	Betty_Currie
	8	Coingate	Monica
	9	Black_Mist_Scandal	Affair
	10	Scandal_(disambiguation)	Breuer
	11	2006_Reuters_fake_photos_ scandal	Charles_Ruff
	12	Boesky_scandal	Robert_S._Bennett
	13	Panama_scandal	Mark_Whitaker
	14	Sex_scandals	David_Horsey
	15	Shell_Scandal_of_1915	1983_congressional_page_sex_scandal

Table 5 below shows the top 10 concepts generated in the semantic profiles for our sample query pairs: “George Bush :: Al Gore” and “Michael Jordan :: Charles Barkley”. For the query pair “Michael Jordan :: Charles Barkley”, the top 10 concepts identified as the interlinking terms using our adapted ESA include their most relevant persons, events, etc., and noise concepts are successfully removed from the semantic

Table 4. Ten query pairs used for the semantic profile generation comparison

Topic A	Topic C
George Bush	Al Gore
Michael Jordan	Charles Barkley
Sadam Hussein	Gulf War
Northern Alliance	European Union
Wall Street	New York Times
Steve Jobs	Mark Zuckerberg
Knowledge Discovery	Document Classification
Abdel Rahman	Blind Sheikh
Saudi Arabia	Kuwait
Terrorist Attack	Bill Clinton

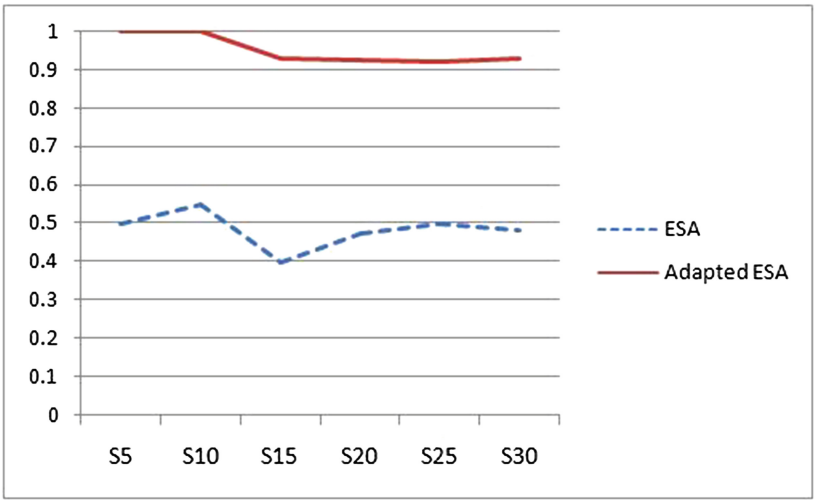

Fig. 8. The Averaged Precision of the Intermediate Semantic Profile (BP profile) Generation using the original ESA and adapted ESA based on processing data in Table 4.

profile and key interlinking concepts are boosted to higher positions such as "I_May_Be_Wrong_but_I_Doubt_It", a memoir by Charles Barkley that recounts some of Barkley's memorable experiences including his involvement with Michael Jordan as a member of the "Dream Team", and "1993_NBA_Finals", the championship round of a historic season when Michael Jordan led the Chicago Bulls to play against the Phoenix Suns which was led by Charles Barkley. By contrast, the original ESA failed to rank high for some very important concepts related to them.

In terms of concept chain queries, we have also conducted a qualitative evaluation of the proposed model for generating various lengths of chains using the precision ratio defined below.

Table 5. Top 10 concepts in the sample semantic profiles generated by the adapted ESA and the original ESA

Input	#	Original ESA	Adapted ESA
George Bush :: Al Gore	1	George_Rose_(disambiguation)	Electoral_history_of_George_W._ Bush
	2	Electoral_history_of_George_ W. _Bush	Al_Gore_presidential_campaign,_ 2000
	3	Sir_Ralph_Gore,_4th_Baronet	Snippy
	4	St_George_Gore-St_George	Non-rigid_designator
	5	Sir_Arthur_Gore,_1st_Baronet	United_States_presidential_election_ in_Massachusetts,_2000
	6	Electoral_history_of_George_H. _W._Bush	High_Performance_Computing_and_Communication_ Act_of_1991
	7	Al_Gore_presidential_campaign, _2000	Millie_(dog)
	8	Tennis_at_the_1908_Summer_ Olympics_– _Men's_indoor_doubles	United_States_presidential_election_in_the_District_ of_Columbia,_ 2000
	9	The_Betrayal_of_America	John_Prescott_Ellis
	10	James_Howard_Gore	George_H._W._Bush
Michael Jordan :: Charles Barkley	1	David_Jordan	I_May_Be_Wrong_but_I_Doubt_It
	2	Charles_Blount	1993_NBA_Finals
	3	Charles_Evans	Barkley,_Shut_Up_and_Jam:_Gaiden
	4	Charles_Bronson_(disambiguation)	Best_NBA_Player_ESPY_Award
	5	Charles_Jordan_(magician)	1996_NBA_All-Star_Game
	6	Baggir	1986–87_NBA_season
	7	Manduca_fosteri	1992–93_NBA_season
	8	I_May_Be_Wrong_but_I_Doubt_It	1990–91_NBA_season
	9	Manduca_diffissa	1991–92_NBA_season
	10	Karl_Jordan	Gaiden

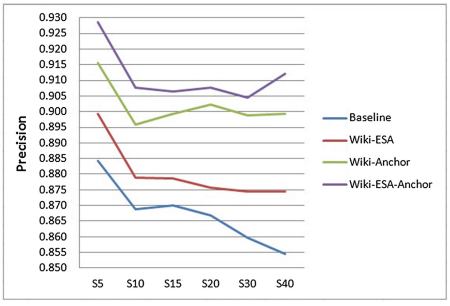

Fig. 9. Search results of chains of length 1

$$precision = \frac{concept\ chains\ found\ and\ correct}{total\ concept\ chains\ found} \tag{12}$$

Figures 9, 10, 11 and 12 make a comparison of the search results in various models. We have implemented a competitive baseline algorithm (i.e. Srinivasan's 'closed' discovery algorithm) where only the corpus-level TFIDF-based statistical information is

considered. In the four figures, the X-axis indicates the number of concepts kept in each semantic type in the search results (S_N means the top N are kept) and the Y-axis indicates the precision values. It is easy to observe that the search performance has been significantly improved with the integration of Wikipedia knowledge, and the best performance is observed when both the Wiki article content and anchor texts are involved.

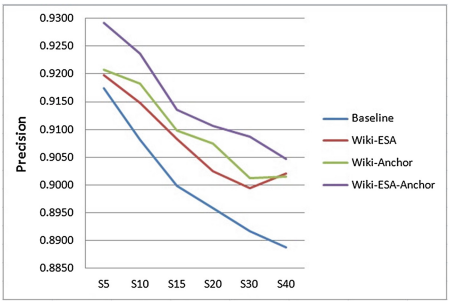

Fig. 10. Search results of chains of length 2

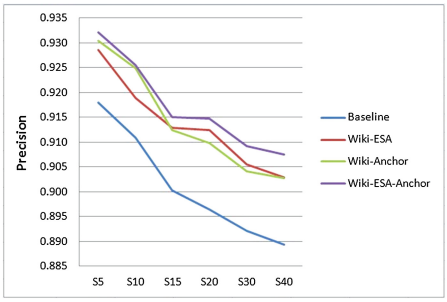

Fig. 11. Search results of chains of length 3

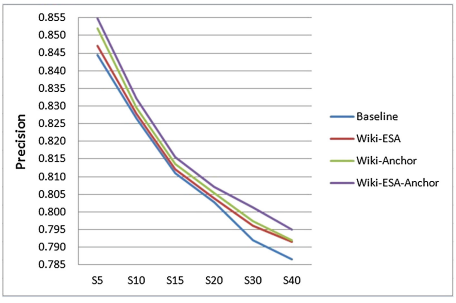

Fig. 12. Search results of chains of length 4

We further used the 37 truth chains described above to measure the performance of the baseline model and various Wiki-enabled models in detecting these chains. In Fig. 13, the X-axis has the same meaning as in Figs. 9, 10, 11 and 12 and the Y-axis now denotes the percentage of the 37 truth chains found by different models. The results also agree with our expectation that the largest percentage of the truth chains were retrieved when incorporating both article content and anchor texts from Wikipedia into the query process.

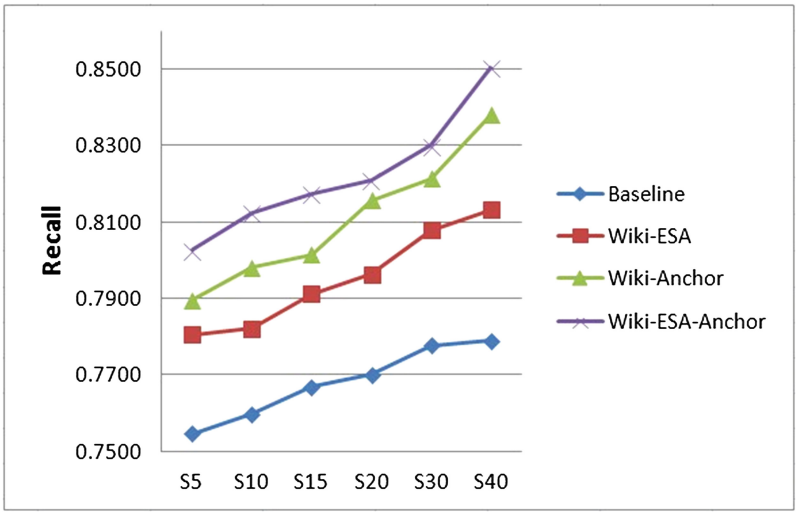

Fig. 13. Comparison of search results using 37 truth chains

Table 6 shows the evidence trails generated for concept chains discovered from Wikipedia. Note that the generated evidence trail is not necessarily from the same Wikipedia article, but could be found through the discovery of knowledge holding across articles. For example, for the concept chain *Betty Ong* → *September 11* → *Mohamed Atta*, sentences were extracted as the supporting evidence from two different Wikipedia articles "*Flight attendant*" and "*American Airlines Flight 11*". Our proposed model successfully found "*Flight attendant*" and "*American Airlines Flight 11*" as two highly relevant Wikipedia articles with regard to "*Betty Ong*" who was a "*Flight attendant*" onboard "*American Airlines Flight 11*" when it was hijacked and flown into the North Tower of the World Trade Center and "*Mohamed Atta*" who was one of the ringleaders of the "*September 11*" attacks, and crashed the "*American Airlines Flight 11*" into the World Trade Center as part of the 9/11 attacks.

Table 6. Evidence trails generated from Wikipedia

Concept chain	Evidence
Betty Ong → September 11 → Mohamed Atta	Sentence 1. The role of flight attendants received heightened prominence after the **September 11** attacks when flight attendants (such as Sandra W. Bradshaw and CeeCee Lyles of United Airlines Flight 93, Robert Fangman of United Airlines Flight 175, Renee May of American Airlines Flight 77 and **Betty Ong** and Madeline Amy Sweeney of American Airlines Flight 11) actively attempted to protect passengers from assault, and also provided vital information to air traffic controllers on the hijackings
	Sentence 2. **Mohamed Atta**, the ringleader of the attacks, and a fellow hijacker, Abdulaziz al-Omari, arrived at Portland International Jetport at 05:41 Eastern Daylight Time on **September 11**, 2001
Rahman → Bin Laden → Al Qaeda	Sentence 1. ***Rahman*** built a strong rapport with ***bin Laden*** during the Soviet war in Afghanistan and following Azzam's murder in 1989 Rahman assumed control of the international jihadists arm of MAK/***Al Qaeda***
Gore → Bush → Stephen Hadley	Sentence 1. ***Bush***, at the advice of ***Hadley***, also proposed greater nuclear arms reductions than ***Gore***
Atta → Huffman → Dekkers	Sentence 1. ***Atta***, along with Marwan al-Shehhi arrived in Venice, Florida, and visited ***Huffman*** Aviation to "check out the facility"
	Sentence 2. On the eve of the trial, ***Dekkers*** sold all of ***Huffman***'s holdings minus 10 planes to Triple Diamond, to gather the money needed to repay his business partner

6 Conclusion and Future Work

This paper proposes a new solution for improving cross-document knowledge discovery through our introduced concept chain queries, which focus on detecting semantic relationships between concepts across documents. In this effort, we attempt to incorporate relevant Wikipedia knowledge into the search process, which effectively complements the existing knowledge in document collections and further improves search quality and coverage. Additionally, a better measure for estimating semantic relatedness between terms is devised through integrating various evidence resources from Wikipedia. Experimental results demonstrate the effectiveness of our proposed new approach and show its advantage of alleviating semantic loss caused by only using the Vector Space Model (VSM) on the corpus level.

Future directions include the exploration of other potential resources provided by Wikipedia to further improve query processing, such as infobox information, categories that relevant Wiki articles belong to and the underlying category hierarchy. These valuable information resources may be combined with our defined semantic types to further contribute to ontology modeling. As a cross language knowledge base, we also

plan to explore the utilization of Wikipedia knowledge in a cross-lingual setting to better serve different query purposes.

References

1. Jin, W., Srihari, R.K.: Knowledge discovery across documents through concept chain queries. In: Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops 2006, pp. 448–452. IEEE, December 2006
2. Srinivasan, P.: Text mining: generating hypotheses from MEDLINE. J. Am. Soc. Inform. Sci. Technol. **55**(5), 396–413 (2004)
3. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611, January 2007
4. Swanson, D.R., Smalheiser, N.R.: Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. Libr. Trends **48**(1), 48–59 (1999)
5. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In: AAAI, vol. 6, pp. 1301–1306, July 2006
6. Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In: Proceedings of the Semantic Web Workshop at SIGIR 2003, November 2003
7. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring web communities from link topology. In: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space—Structure in Hypermedia Systems: Links, Objects, Time and Space—Structure in Hypermedia Systems, pp. 225–234. ACM, May 1998
8. Sorg, P., Cimiano, P.: Cross-lingual information retrieval with explicit semantic analysis. In: CLEF Workshop 2008 (2008)
9. Scott, S., Matwin, S.: Text classification using WordNet hypernyms. In: Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference, pp. 38–44, August 1998
10. Srihari, R.K., Li, W., Niu, C., Cornell, T.: Infoxtract: a customizable intermediate level information extraction engine. In: Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems, vol. 8, pp. 51–58. Association for Computational Linguistics, May 2003
11. Faloutsos, C., McCurley, K.S., Tomkins, A.: Fast discovery of connection subgraphs. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 118–127. ACM, August 2004
12. MWDumper. Software available at http://www.mediawiki.org/wiki/Manual:MWDumper
13. Yan, P., Jin, W.: Improving cross-document knowledge discovery using explicit semantic analysis. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2012. LNCS, vol. 7448, pp. 378–389. Springer, Heidelberg (2012)
14. Jin, W., Srihari, R., Singh, A.: Generating hypotheses from the web. In: Proceedings of the 17th International Conference on World Wide Web, pp. 1211–1212. ACM, April 2008
15. Luo, G., Tang, C., Tian, Y.L.: Answering relationship queries on the web. In: Proceedings of the 16th International Conference on World Wide Web, pp. 561–570. ACM, May 2007
16. Radev, D.R., Libner, K., Fan, W.: Getting answers to natural language questions on the Web. J. Am. Soc. Inform. Sci. Technol. **53**(5), 359–364 (2002)
17. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. In: WWW 2007, pp. 757–766 (2007)
18. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)

19. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. arXiv preprint cmp-lg/9808002 (1998)
20. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, pp. 519–528. ACM, May 2003
21. Jing, L., Zhou, L., Ng, M.K., Huang, J.Z.: Ontology-based distance measure for text clustering. In: Proceedings of the Text Mining Workshop, SIAM International Conference on Data Mining (2006)
22. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
23. Rodríguez, M.D.B., Hidalgo, J.M.G., Agudo, B.D.: Using WordNet to complement training information in text categorization. arXiv preprint cmp-lg/9709007 (1997)
24. Gurevych, I., Müller, C., Zesch, T.: What to be?-electronic career guidance based on semantic relatedness. In: Annual Meeting-Association for Computational Linguistics, vol. 45(1), p. 1032, June 2007
25. Müller, C., Gurevych, I.: Using wikipedia and wiktionary in domain-specific information retrieval. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 219–226. Springer, Heidelberg (2009)
26. Jin, W., Srihari, R.K., Ho, H.H., Wu, X.: Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In: Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 193–202. IEEE, October 2007
27. Yan, P., Jin, W.: Mining semantic relationships between concepts across documents incorporating wikipedia knowledge. In: Perner, P. (ed.) ICDM 2013. LNCS, vol. 7987, pp. 70–84. Springer, Heidelberg (2013)
28. Bonifati, A., Cuzzocrea, A.: Efficient fragmentation of large XML documents. In: Wagner, R., Revell, N., Pernul, G. (eds.) DEXA 2007. LNCS, vol. 4653, pp. 539–550. Springer, Heidelberg (2007)
29. Cuzzocrea, A., Darmont, J., Mahboubi, H.: Fragmenting very large xml data warehouses via k-means clustering algorithm. Int. J. Bus. Intell. Data Min. **4**(3), 301–328 (2009)
30. Cuzzocrea, A., Bertino, E.: A secure multiparty computation privacy preserving OLAP framework over distributed XML data. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1666–1673. ACM (2010)
31. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**(1), 141–188 (2010)
32. Deerwester, S.: Improving information retrieval with latent semantic indexing. In: Proceedings of the 51st Annual Meeting of the American Society for Information Science, pp. 36–40 (1988)
33. Meng, L., Huang, R., Gu, J.: A review of semantic similarity measures in wordnet. Int. J. Hybrid Inform. Technol. **6**(1), 1–12 (2013)
34. Rinaldi, A.M.: An ontology-driven approach for semantic information retrieval on the web. ACM Trans. Internet Technol. (TOIT) **9**(3), 10 (2009)
35. Wu, H., Gunopulos, D.: Evaluating the utility of statistical phrases and latent semantic indexing for text classification. In: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2003, pp. 713–716. IEEE (2002)
36. Liu, T., Chen, Z., Zhang, B., Ma, W.Y., Wu, G.: Improving text classification using local latent semantic indexing. In: Fourth IEEE International Conference on Data Mining, ICDM 2004, pp. 162–169. IEEE, November 2004
37. Salahli, M.A.: An approach for measuring semantic relatedness between words via related terms. Math. Comput. Appl. **14**(1), 55 (2009)