2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

Revisiting Concepts of Topicality and Novelty

- A New Simple Graph Model that Rewards and Penalizes Based On Semantic Links

Fei Yu, Eugene Santos, Jr.
Thayer School of Engineering
Dartmouth College
Hanover, N.H., U.S.A
{fei.yu, eugene.santos.jr}@dartmouth.edu

Abstract—**Research in Information Retrieval (IR) experienced a paradigm shift from first having too few documents to search from to now having way too many of them. When users have trouble finding relevant documents, they tend to become frustrated and give up searching. Scholars have attempted to reduce instances of search frustration via query expansion, information filtering, and incorporating user feedback. However, these approaches are not effective as users still experience a period of frustration before getting more relevant results. The aim of this conceptual paper is to explore possible improvements to the field by revisiting two fundamental concepts: topicality and novelty. First, we elaborate various issues with existing IR models in capturing these two concepts. Second, we illustrate a potential improvement to these issues: namely, a new simple graph space model with new topicality and novelty measures that can better capture these features of a document based on rewards and penalties for corresponding matching and missing semantic links. Lastly, we demonstrate a walk-through example using the new graph-based IR model.**

Keywords-information retrieval; graph space model; vector space model; topicality; novelty; diversity

I. Introduction

The ultimate goal of an information retrieval (IR) system is to select information and organize them in the same way as a real human does. The most critical step in building a successful IR system is to simulate how a real human judges the relevancy of each document. *Human-centered research* focuses on the correspondence between one's underlying cognitive/psychological processes and his relevance judgment of a document. On the other hand, *system-centered research* focuses on learning the correspondence between relevance judgment and the query-document pair. Since human-centered research studies how humans think whereas system-centered research examines how humans behave, it is challenging to bridge the research findings from one stream to the other.

Based on the results from various empirical and exploratory studies, human-centered research suggests that the relevance judgment of a document should be subjective, multidimensional, and dynamic [1]. However, it is challenging for an IR model to capture these notions since there is still a lack of understanding what subjective, multidimensional, and dynamic relevance constitutes. Out of all possible criteria that are measurable from queries and documents, the *topicality* and *novelty* criteria have received the most interest from both research streams [2][3]. The topicality of a document is concerned with the question of whether the topic of a document matches the topic of a user's information need. It has so far been the most prevalent measure of relevance. The novelty criterion is concerned with one's perception of a document based on the information he already knows about. If there are two nearly duplicate documents, a user may be only interested in reading one of them. Even though these two documents are similar in terms of topicality, the document picked first might be considered more relevant than the second one. As a consequence, novelty is a dynamic feature of a document that changes over time according to the documents accessed by users.

The aim of this conceptual paper is to revisit the two fundamental concepts of topicality and novelty. As generative models tackle the IR problem from the perspective of likelihood, our discussions in this paper focus on discriminative models that are originated to match topicality between document/query pair. In particular, it has been identified that current efforts of discriminative models in topicality fall short of the target human notions. This is a similar problem for novelty which is often over simplified. Even more challenging, novelty and topicality are typically cross-defined and interdependent. The problem gets further worsened by the overwhelming number of documents ever available due to the wide use of world wide web. Very often, out of all the documents returned by a search, only a small number of them are topically relevant. Studies on user frustration experience [4][5] have been reported and the source of frustration has been investigated. Scholars have attempted to reduce instances of search frustration via query expansion [6], information filtering [7], and incorporating user feedback [8]. However, the problem of capturing the notion of topicality and novelty still remains unsolved. The word independence property of vector space models (e.g. term frequency-inverse document frequency models [9]) has long been argued to be the one of the culprits

978-1-4673-1714-6/12/$31.00 ©2012 IEEE

of poor retrieval performance. However, little consensus has been reached so far on how to break this word independence assumption. If an IR model considers modeling dependency relations between two words, what type of dependencies (e.g. syntactic dependence, word co-occurrence, discourse relations etc.) is the right one to capture? Furthermore, how can an IR model improve the topicality and novelty measures via identifying these relations? As natural languages are flexible and diverse in nature, an IR model may also be error prone in comparing these relations across documents.

In order to improve the existing topicality measure, we illustrate one potential improvement: a new simple graph space model that can better capture topicality and novelty features of a document. This graph space model breaks the word independence assumption by proposing two types of dependency relations: *contextual relations* and *lexical relations*. Simply put, contextual relations link concepts that are talked about together while lexical relations link concepts that have same/similar semantic meanings. When computing the topicality measure, contextual relations help to define a user's information request while lexical relations help to relax a user's information request appropriately to avoid mismatching with a document. In addition to the new topicality measure based on our graph space model, we also propose a novelty measure that can better discover novel content within a document.

This paper is organized as follows: In Section II, we discuss how IR models interpret the notion of topicality and novelty. In Section III, we describe the construction of a graph representation and appropriate algorithms to compute topicality and novelty. In Section IV, we walk through an example and compare results of our topicality measure against the classical similarity-based IR models. Conclusions and future directions can be found in Section V.

II. Related work

This section is a survey of existing IR models with a focus on the relationships between topicality and novelty. The notion of novelty and diversity are often used interchangeably in prior work. In this paper, we define diversity as a spectrum that can be as broad as referring to a piece of new information that a user has not read or as specific as referring to a new subtopic or a new argument. One end of the diversity spectrum is called novelty, which refers to information about new subtopics that a user has not read. The other end of the diversity spectrum is low redundancy, which refers to information that has little overlapped content with others. As topicality describes how well a document matches a user's information need while diversity deals with how well a document differs from a user's knowledge state, the notion of diversity does not subsume the notion of topicality. The human-centered research considers the concept of novelty to be complementary with the concept of topicality. However, some IR models treat diversity as a compensation for inaccurate topicality measures rather than as a complement. In this section, we categorize IR models by their relations between novelty and topicality measures.

A. Diversity measure as compensation to topicality measure

The topicality measurement of an IR model relies heavily on the representation of information needs and each document. As illustrated by Sandor Dominich [10], a user first realizes an information need in his mind (e.g., wants to know about a specific topic, looking for a book to study a subject, etc.). Second, he translates his information need into an information request in natural language. Finally, he translates his information request into a query which consists of word units. The first category of diversity-seeking approaches address what is missing during these two translations.

1) Missing information from translating an information need to an information request

When an information request is formed, only part of the information need is covered in the request. The information request focuses on describing what the targeted information is about rather than clarify what it is not about. Furthermore, a user is often not aware of multiple possible interpretations of his own information request. The missing information, even though they are not as useful in describing one's information needs, may play a critical role in resolving ambiguities in the information request. In order to mitigate the situation where the topicality measurement fails to distinguish among multiple possible interpretations, some diversity-seeking IR models [11] deliberately look for results that cover different interpretations.

2) Missing information from translating an information request to a query

An information request is expressed in natural language, e.g., "How to fix a coffee machine with model number XB360?". A user needs to translate a request into a keyword-based query, in a form that can be understood by a typical IR system. For example, the request above can be translated to a query such as "coffee machine model XB360 repair" or "coffee maker model XB360 fix". The order of words within this query does not make any difference to an IR system based on the vector space representation. Thus, a lot of semantic meaning is lost when a query is represented via a vector, which is the key input to measure topicality. Even though not explicitly stated, the diversity measure of IR models [12] help promote diverse information that previous documents do not contain with the hopes that some can satisfy the user. Back to the coffee machine example, a matched document may be mainly about how to fix a model XB260 rather than a model XB360 but only briefly mentioned model XB360. This document would be considered highly topical since it covers all the keywords (coffee, machine, model, XB360, repair) of the query. However, the document will not be considered to be of much use by the user. By keep retrieving diverse documents, a relevant document talking about repairing a model XB360 may get retrieved.

B. Diversity measure as complement to topicality measurement

Different from the first category, the second category of diversity-seeking approaches assumes that the topicality measurement is a sound simulation of human judgment. This category includes approaches that supply documents covering diverse sub-topics and supply documents containing little overlapping information among each other.

1) Diversity-seeking to discriminate documents that have high redundancy

Maximal Marginal Relevance (MMR) models and its variants are popular approaches to retrieve documents with minimal overlapped content [12]. The diversity measure is captured by computing how dissimilar its content is compared against the documents with higher ranks. As a consequence, documents with high redundancy are discriminated by MMR models. There are two issues with the MMR models. First, documents with low redundancy do not equate with higher relevancy. Second, the rank of a document containing sections of high topicality may get compromised if it also contains sections of high redundancy.

2) Diversity-seeking to discover novel content

Recently, various approaches aim to supply results that cover diverse subtopics. Carterette et al. [13] propose promoting documents with new subtopics. One question that remains unanswered is whether novel content is a strong indicator of relevancy. In other words, if two documents are both topical to a user's query, is the more novel document more relevant? Xu and Yin [14] state that novelty seeking is not equal to diversity seeking. They propose providing novel documents by directing a user to a certain subtopic area, which is referred to as a directed novelty-seeking approach. They suggest that novel content is potentially more relevant if its topics are among a user's interests.

C. Conclusions

We summarize the main issues in capturing topicality and novelty as follows:

(1) The current topicality measure is an imperfect simulation of a real human's topicality judgment.

(2) The current novelty measure suffers from inaccurate topicality measurements.

(3) The current novelty measure only captures one type of diversity, although information is diverse in multiple aspects.

Unfortunately, as the information space we are dealing with nowadays is of massive volume, highly redundant, and dynamic in nature, current IR models are far from being satisfactory. We believe that a more accurate topicality measure would have large impacts on the definition and modeling of novelty, and most importantly to a sound relevancy measure.

III. Graph representation

In order to better capture the notion of topicality, we illustrate a simple graph model that focuses on the semantic meaning of a document. The graph model relies on two types of dependency relations: *contextual relations* and *lexical relations*. Simply put, contextual relations link concepts that are talked about together while lexical relations link concepts that have same/similar semantic meanings. When computing the topicality measure, contextual relations penalizes the score of a document if relations expected by an information request are not found in a document. On the other hand, lexical relations rewards the score to concepts found in a document with similar semantic meanings to the ones in an information request. In the past, this graph representation has shown promising performance in solving tasks in different domains such as user modeling [15], insider threat detection [16], and cognitive style classification [17]. Within this section, we first describe how a graph representation is constructed with some detailed justification of its advantages over traditional vector space representations. Second, we propose some principles to capture topicality of a document using graph representations. Then, we illustrate how each type of diversity can be individually captured.

A. Graph Representation Construction

Converting a natural language text to a graph representation is a pipeline process as illustrated in Fig. 1. The pipeline process splits a natural text into a list of sentences (sentence segmentation), generates a constituent tree using a parser such as Link Grammar [17] from each sentence (link grammar parsing), extracts semantic links between noun phrases based on four pre-defined heuristics (link extraction), remove stop words in noun phrases (e.g. a, the, his etc.) and conduct word stemming , and finally, build a graph based on semantic links (graph construction). We now demonstrate the conversion process of the following sentence: “Before making a house purchase, a family needs to decide their budget first.” to a graph representation in Fig. 2.

Nodes within the graph representation are called concept nodes, representing the main entities of a sentence. Contextual relations (produced by Sentence-heuristic, Prepositional Phrase-heuristic, and Cross Sentence-heuristic) specify the contextual information of a concept. Sentence-heuristic generates a relation between two noun phrases that are connected by a verb phrase. The relation between concept *family* and concept *budget* is an example of a contextual relation generated by the Sentence-heuristic (Fig. 2(a)). Prepositional Phrase-heuristic recognizes relations between two noun phrases connected by a prepositional phrase. Cross Sentence-heuristic is a (new) heuristic that looks for contextual relation between clauses and between sentences. The relation between concept *house purchase* and concept *family* is an example of a contextual relation generated by Cross Sentence-heuristic. Lexical relations (produced by Noun Phrase-heuristic) identify concepts with identical or similar meaning. Within a noun phrase, the heuristic identifies a *set-subset* relation. The relation between concept *house purchase* and *purchase* and the relation between *house purchase* and *house* are examples of relations generated by Noun-Phrase heuristic (Fig. 2(a)). The intuition behind the Noun Phrase-heuristic is that a concept can be referred to with its modifier omitted. For instance, a sentence may use word *purchase* to refer to a *house purchase* if the complete noun phrase is already mentioned in previous sentences.

From a performance perspective, all the relations are extracted based on noun phrases due to two reasons: First, the graph representation remains robust to errors produced by syntactic parsers (Link Grammar). Second, it is less computationally expensive to conduct shallow linguistic analysis via heuristics than to conduct deep linguistic analysis.

B. Advantages of Graph Representations in Capturing Topicality

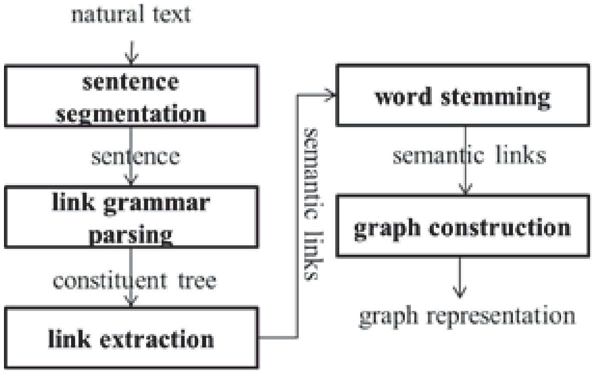
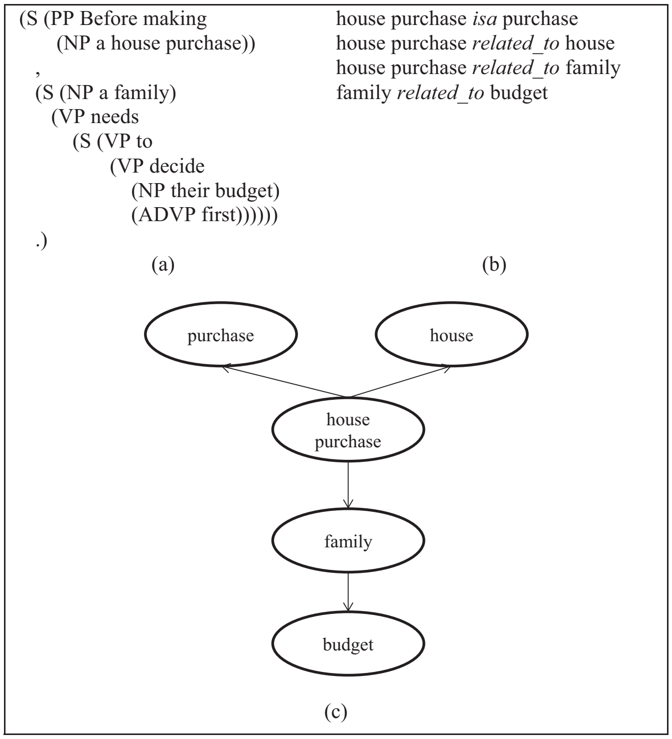


Figure 1: Process of converting a natural text to a graph representation



```
(S (PP Before making
      (NP a house purchase))
  ,
  (S (NP a family)
     (VP needs
        (S (VP to
              (VP decide
                 (NP their budget)
                 (ADVP first))))))
  .)
```

(a)

house purchase *isa* purchase
house purchase *related_to* house
house purchase *related_to* family
family *related_to* budget

(b)

Figure 2: An Example of Converting a Text to a Graph Representation (a) Constituent Tree generated by Link Grammar [17] (b) Semantic Relations (c) Graph Representation

Computational models based on vector space representations capture the notion of topicality as a similarity value between a query and a document. Here, we illustrate the impacts of the word independence assumption on a query and on a candidate document using examples. With regards to queries, the query *house purchase* and *purchase house* are treated identically by a vector-based IR system. However, the first query is more likely to refer to a recent house purchase while a user who issued the second query may still be in the process of house hunting. The information needs expressed by these two queries are not differentiable by vector space representations. Research on query modifications is motivated towards distinguishing different information needs behind these two queries, given some understanding of the user. In terms of a candidate document, a document that talks about *home decoration* is likely to mention both the words *purchase* and *house*. However, the word *purchase* may refer to purchasing furniture, paint etc. rather than referring to a purchase event of a house. In this case, a document may be mistakenly considered to be highly topical. In summary, computational models based on vector space representations are insufficient for expressing a user's topicality request as well as insufficient in expressing the information conveyed in each individual document.

We believe the new graph model to be more effective in expressing both a user's request and information within individual documents. Advances in natural language processing research allow an IR system to directly interpret the information request instead of asking a user to transform his information needs into a query. Similarly, since each document is also represented in the form of a graph, it allows more accurate matching between a document and information request. An example showing how different sentences are converted to a graphical representation can be found in Fig. 3. The two sentences have the same vector space representations[1] with a TFIDF weighting scheme excluding the stop words (a possible TFIDF scheme of these two sentences is shown in the rightmost column). The graph representations of these two sentences capture their different semantic meanings: the concept of *family* is related to the concept of *house purchase* in the 1st sentence while is related to the concept of *furniture purchase* in the 2nd one.

A graphical representation, as we can observe from the above example, can better capture the semantic meaning behind one's information request by explicitly requiring semantic relations. A potential problem of a graphical representation is that the semantic relations can be too restrictive to match a document. We now discuss this issue at both the sentence level and document level. On the sentence level, the same semantic meaning can be delivered using sentences with different grammar structures. On the document level, the same semantic meaning can be conveyed by sentences organized in different ways. Therefore, the graph representation we propose only involves nouns/noun phrases as they are relatively static words. Contextual links are relatively insensitive to different surface structures if their underlying deep structures are the same [19]. For example, the following three sentences (s_2 and s_3 are paraphrases of s_1) have the same graph representations.

s_1: Before making a house purchase that includes furniture, a family needs to decide their budget first.

s_2: A family needs to decide their budget before making a house purchase with furniture.

s_3: A budget needs to be decided by a family before they make a house purchase that covers furniture.

[1] . TFIDF weighting scheme is used to construct the vector space representation for two sentences with the following stop words excluded: before, make, a, for, with, need, to, their, first. All the inverted document frequency values are synthesized rather than derived from a real corpus as the idf for each term is static.

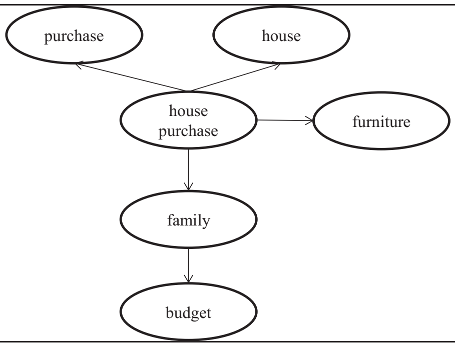
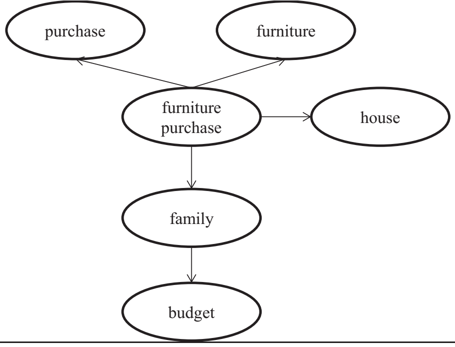
<table>
<tr><th>Sentence</th><th>Graph Representation</th><th>Vector Space Representation</th></tr>
<tr><td>Before making a house purchase that includes furniture, a family needs to decide their budget first.</td><td> </td><td>d₁=[0.667,0.2554,0.6234,0.5860,0.3343,2.333]<table>
<tr><td>Index</td><td>Term</td><td>tfidf weight</td></tr>
<tr><td>1</td><td>house</td><td>0.6667</td></tr>
<tr><td>2</td><td>purchase</td><td>0.2554</td></tr>
<tr><td>3</td><td>furniture</td><td>0.6234</td></tr>
<tr><td>4</td><td>family</td><td>0.5860</td></tr>
<tr><td>5</td><td>decide</td><td>0.3343</td></tr>
<tr><td>6</td><td>budget</td><td>2.3333</td></tr>
</table></td></tr>
<tr><td>Before making a furniture purchase for a house, a family needs to decide their budget first.</td><td> </td><td>d₂=[0.667,0.2554,0.6234,0.5860,0.3343,2.333]<table>
<tr><td>Index</td><td>Term</td><td>tfidf weight</td></tr>
<tr><td>1</td><td>house</td><td>0.6667</td></tr>
<tr><td>2</td><td>purchase</td><td>0.2554</td></tr>
<tr><td>3</td><td>furniture</td><td>0.6234</td></tr>
<tr><td>4</td><td>family</td><td>0.5860</td></tr>
<tr><td>5</td><td>decide</td><td>0.3343</td></tr>
<tr><td>6</td><td>budget</td><td>2.3333</td></tr>
</table></td></tr>
</table>

Figure 3. Two Sentences with Their Respective Graph and Vector Representation

C. Topicality Measure

Based on the graph model, we describe a new topicality measure derived in two steps based on our graph representation.

$$Topicality(Q,DG) = (\alpha+1)\cdot closeness(V') \qquad (1)$$

where Q represents an information request and DG represents a document, both of which are graph representations. V' is the set of topical concepts of a document, and α is a penalty on indirect topical mappings computed for Step 1. In Step 2, closeness(V') computes how far on average a topical concept is from another one within a DG.

Step 1: identify topical concepts

This step maps topical concepts of an information request with their semantic equivalents in each document. The topical matching process is a recursive process to match all noun phrases of an information request (pseudo code is presented Fig. 4). If a noun phrase is not found, the matching process considers concepts with lexical relations (we call such concepts as *substitutes*). For example, if a document does not have the noun phrase *longest single-span bridge*, the matching process looks for its substitutes: *longest bridge*, *single-span bridge, longest, single-span,* and *bridge*. If any substitute is not found, the process continues to check for its substitutes. For example, if *single-span bridge* is not found, the matching process continues to seek *single-span* and *bridge* respectively. The number of operations α refers to the total number of substitutions that occurs in the matching process. In this paper, the substitutes of a noun phrase are generated using the NP-heuristic. More sophisticated linguistic processes such as synonyms, hypernyms, and hyponyms identifications can also help seek concepts with similar semantic meaning.

Step 2: compute closeness of topical concepts

Once topical concepts are identified within a document, Step 2 computes how close these topical concepts are related to each other. The intuition behind this measure is that the closer all these topical concepts relate to each other, the more likely these topical concepts are all describing the information that a user is targeting at. The function to compute closeness is presented in Equation (2).

$$closeness(V') = \frac{2\sum_{v_i \in V'}\sum_{V_j \in V', j \neq i} shortestpath(V_i', V_j')}{|V'|\cdot(|V'|-1)} \qquad (2)$$

where V' is the set of topical concepts identified in Step 1. The shortest path (in this paper we use Floyd-Warshall algorithm) between two topical concepts is computed on the graph representation of the entire document. Since the closeness measure denotes how far on average one topical concept is from another one, the measure is normalized by the total number of shortest paths. It is guaranteed that a shortest path between two concepts can be found because of the Cross Sentence-heuristic.

D. Novelty Measure

Traditional information-retrieval approaches consider topicality and novelty independent of each other. Maximal Marginal Relevance (MMR) is a popular diversity-seeking approach.

$$MMR(D_i) = Arg\max_{D_i \in R \backslash S}(\lambda Sim(D_i,Q) + (1-\lambda)\max_{D_j \in S}(Sim(D_i,D_j))) \qquad (3)$$

D_j is a document belonging to the set of all documents ($D_j \in R$) and is already selected ($D_j \in S$), Q is user query/quest, and D_i is a document that has not been selected yet ($D_i \in R/S$). λ is a weight to balance topicality and diversity measures.

The topicality and diversity metrics computed in MMR are treated as two compensatory components. The intuition behind

MMR is that a low-topical document may still be relevant to a user if it contains novel content. On the other hand, a high-topical document can be of little use if it provides little new information. The approach sounds intuitive, but it may not perform empirically well due to two issues. First, both the topicality and diversity metrics are based on entire document D_i. Thus, the diversity metric of MMR reflects how different the entire document is from a user's knowledge state rather than consider the content closely associated with a user's information request. As vector space representations lose information in terms of how closely other words associate with the words in information request, it is incapable of identifying nearby-concepts of the information requests. The second limitation is its dissimilarity component (diversity metric). A document may be considered diverse due to its lexical differences rather than information novelty. Here, lexical differences refer to different choices of lexicons, anaphora that conveys same/similar semantic meaning. Information novelty refers to a piece of information that may alter a user's knowledge status. Intuitively, a computational model should be able to identify and separate out lexical differences so that the diversity measure can truly represent the degree of content novelty.

In this paper, we also describe a novelty-driven model to discover novel information that is closely related to its corresponding topicality. To be more specific, our model focuses on contextual information surrounding each topical concept. We call our approach *topicality diversification*.

```
TOPICAL-CONCEPTS-MATCHING

Let S be a stack of set {np_i}, each np is a noun phrase of
the information request
Let DG = (V, E)
Let V' be an empty set
Let α be the number of operations requires to finish
topical c ncepts matching
function fi dTopicalC ncepts (S, DG)
    while S is not empty
        e ← pop(S)
        if e ∈ V
            add e to  ’
        else
            produce a stack S’ co taining substitutes of e
             α ← α + 1
             findTopicalConcepts(S’,  G)
```

Figure 4. Pseudo Code for Matching Topical Concepts

$$Novelty(D_i) = \frac{1}{\max\limits_{D_j \in S_k}(Sim_r(SG_i, SG_j))} \tag{4}$$

The novelty metric of a document D_i is computed as the reciprocal value of the largest similarity measure between its *diversified topicality graph* SG_i (pseudo code is presented Fig. 5) and the graph SG_j of other documents (D_j) that are already selected (referred to as a set S). The similarity measure between two graphs focuses on contextual relations rather than lexical relations.

```
  TOPICALITY-DIVERSIFICATION-CONSTRUCTION
1 Let G’ = (V, E) where both V, E are empty
2 Let V' be the set of topical concepts identified in step 1
3 function constructTopicalityDiversifiedGraph (V’, DG)
4     for all v ∈ V’
5         add v, all nodes connected to v in DG, and corresponding
6          edges to G’
```

Figure 5. Pseudo Code for Constructing Diversified Topicality Graph

IV. EXAMPLE

In order to determine whether the new topicality and novelty measures improve retrieval performance, it is best to first illustrate with an example rather than conduct a large-scale empirical test. We walk through an information retrieval task where a user is seeking information about the longest bridge in London. We choose this task because semantic relations in the information request and those in a candidate document play critical roles in retrieval. With this example, we aim to show the importance of breaking word independence and also demonstrate how the topicality of a query gets captured in the new graph representation. In terms of novelty metrics, we demonstrate how diversified topicality graphs look like for several documents.

A. *An IR Task*

Imagine a user realizing an information need to seek information about the longest bridge in London. The first step he would do is to find out the name of the longest bridge in London. We assembled eleven documents from the internet that include relevant documents, irrelevant documents containing incorrect content, and irrelevant documents with missing content, most of which contain the keywords: *longest*, *bridge*, and *London.* His information request is the sentence: “What is the longest bridge in London?” Among all of the eleven documents, only d_9 (description is bolded in the following list) contains the information that the user is looking for. Even though these documents all contain keywords that seem topical to the information request, documents d_1, d_2, and d_3 talk about the longest bridges in other cities other than ones in London. Documents d_4–d_8 talk about the longest span of a bridge rather than describe a bridge to be the longest.

Brief descriptions of the eleven documents:

d_1: the longest single-span bridge in England (Humber Bridge) and famous bridges in London (London Bridge, Albert Bridge)

d_2: the longest single-span bridge in England when it was built (Galton Bridge) and famous bridges in London (London Bridge, Albert Bridge)

d_3: longest single-span suspension bridge in the world when it was built (Tower Bridge)

d_4: a famous bridge in London with a description of its present total length and longest span (Albert Bridge)

d_5: a famous bridge in London with a description of its present total length (Blackfriars Bridge)

d_6: a famous bridge in London with a description of its longest spans and total length but does not mention its design type (Millennium Bridge)

d_7: a famous bridge in London with a description of its longest span and total length as well as its design type (Millennium Bridge)

d_8: a famous bridge in London with a description of its longest span (Waterloo Bridge)

d_9: a famous bridge in London with a description of its longest span and total length. It explicitly states that the Waterloo Bridge is the longest bridge in London. (Waterloo Bridge)

d_{10}: a list of famous bridges in London

d_{11}: a list of different types of bridges (arch bridge and suspension bridge)

We compare the rank produced by our IR model against the one produced by a similarity-based IR model (cosine similarities) based on vector space representation (TFIDF weighting scheme). According to Table I, documents that contain noun phrases *longest bridge* and *London* (d_1, d_2, and d_9) are assigned with highest ranks by our IR model compared to other documents containing their substitutes. However, similarity-based IR model assigns highest ranks to d_6, d_7, and d_8, all of which describe the length of a bridge's longest span. In addition, the best-matching document d_9 only ranks 7th by the similarity-based IR model. Both IR models assign lowest ranks to d_3 and d_{11} as neither of them contains noun phrase *London*.

TABLE I. COMPARISONS OF TOPICALITY RANKS

Document	Topicality Rank (Graph Representation)	Topicality Rank (Vector Space Representation)
1	2	4
2	3	5
3	10	10
4	5	8
5	7	9
6	8	3
7	9	2
8	6	1
9	**1**	**7**
10	4	6
11	11	11

In terms of a novelty measure, let us look at the diversified topicality graphs for d_1, d_2, d_3, and d_9 in Fig. 6. Out of four documents, d_2 and d_9 contain novel content on the topical concept *longest bridge*. This topical concept in d_2 is related to a subtopic about *second-seven crossing* while in d_9 it is related to a subtopic about *Waterloo bridge*. These graphs are also intuitive in terms of explaining why these documents are not the best topical-matching documents. In d_1, the *longest bridge* being described is the *longest single-span bridge* while the *longest bridge* described in d_3 is *longest single-span suspension bridge*. Since being the longest single-span bridge is an indicator of being the longest bridge in the city, concepts with lexical relations tells additional information on these topical concepts.

V. CONCLUSION AND FUTURE WORK

Topicality and diversity are the two most prevalent concepts that reflect a document's relevancy. In this paper, we first pointed out inadequacies of current IR models, particularly vector space models as a classical example, in capturing the semantic meaning of information requests and candidate documents. Consequently, IR models based on vector space models result in inaccurate measurement of a document's topicality. In addition, vector space models limit approaches that aim to discover novel information. As such, we proposed an approach based on a new simple graph space representation. The basic idea is to penalize if expected semantic links are not found in a document and reward if similar concepts (even though not identical) are found in a document. In other words, the contextual relations prevent documents with the same wording but incorrect semantic meaning from being matched while the lexical relations encourage matching documents with inexact wording but similar semantic meaning, which naturally overcome the inadequacies of vector space models. Since novelty is dynamic and unique to each user's knowledge state, there are no readily available/appropriate testbeds. Furthermore, it helps to understand the advantages of new topicality and novelty measure via an example. As such, we have initially demonstrated our concept by walking through an example of a common IR task. Our walk-through shows that a graphical representation is more accurate in capturing topicality and also provides richer information for elaborating the kind and scope of diverse information.

In the future, we will carefully evaluate whether this IR model can perform robustly on large-scale corpora constructed with respect to the different facets (e.g., user-centrism) of novelty. We also realize that it is more computationally complicated to build a graph space representation than to build a vector space representation. Since both our topicality and novelty measure only involve a small number of concepts rather than the entire document, we will investigate techniques to build graph representations that only include these concepts. Lastly, due to limitations of current syntactic parsers especially in noun phrase identification, we aim to continue exploring methods to improve our heuristics in generating contextual and lexical relations.

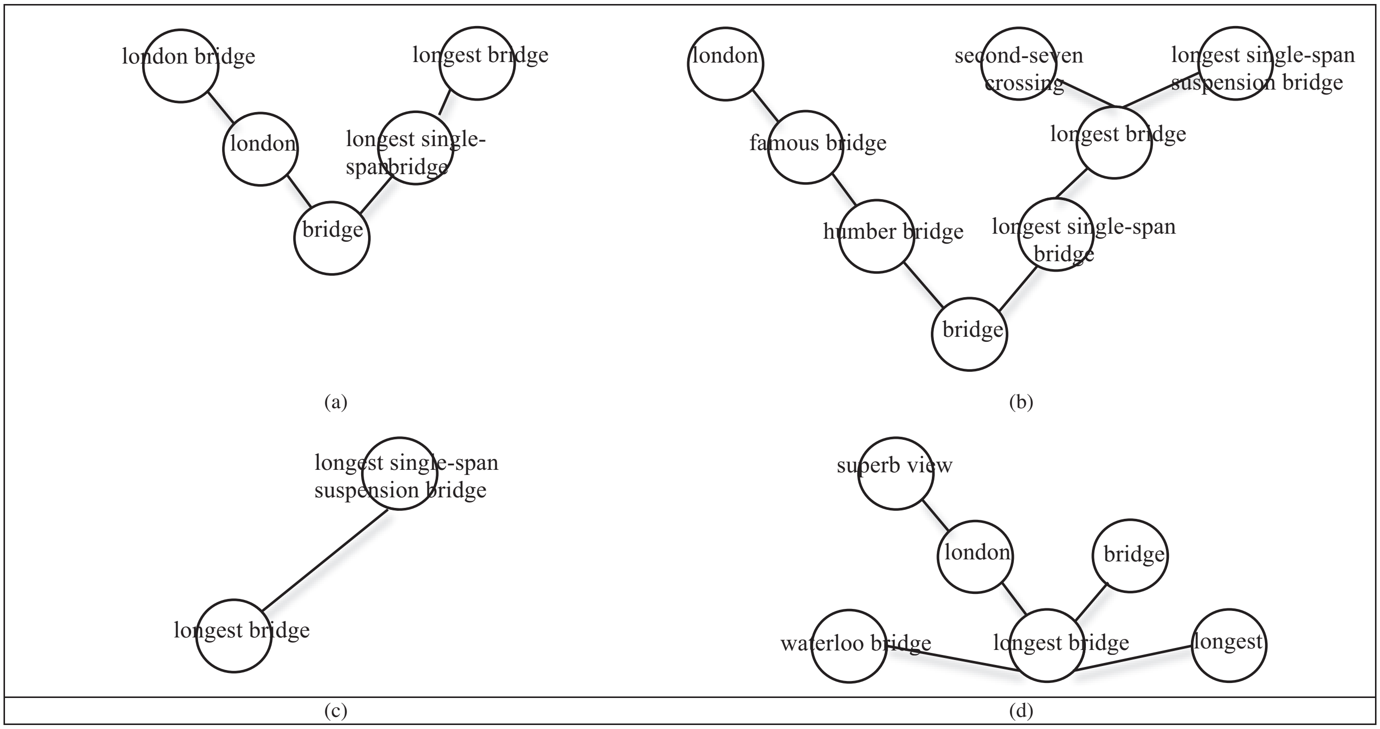


Figure 6. Examples of Diversified Topicality Graphs

Diverisified Topicality Graph for (a) d1 (b) d2 (c) d3 (d) d9

ACKNOWLEDGMENT

This work was supported in part by grants from the Air Force Office of Scientific Research and the Office of Naval Research. We also wish to thank Shiyuan Liu for her help in the literature review.

REFERENCES

[1] W. S. Cooper, “A definition of relevance for information retrieval,” Information Storage and Retrieval, vol. 7, Jun. 1971, pp. 19-37.

[2] B. Boyce, "Beyond topicality:: A two stage view of relevance and the retrieval process," Information Processing & Management, vol. 18, no. 3, pp. 105-109, 1982.

[3] C. Zhai, W.W. Cohen, and J. Lafferty, “Beyond independent relevance: methods and evaluation metrics for subtopic retrieval,” Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, p. 10–17.

[4] Feild, H., Allan, J., & Jones, R., “Predicting searcher frustration”. *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010,* pp. 34-41.

[5] Lazar, J., Bessiere, K., Ceaparu, I., Robinson, J., & Ben, S,, “Help! I’m Lost: User Frustration in Web Navigation”. *It&Society*, *vol. 3*, 2003, pp. 18-26.

[6] Voorhees, M., E., “Query expansion using lexical-semantic relations,” Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994, pp. 61-69.

[7] Belkin, N.J. and Croft, W.B., “Information filtering and information retrieval: two sides of the same coin?” Journal of the Communications of the ACM, vol. 35, 1992, pp. 29-38.

[8] Wang, X. and Fang, H. and Zhai, C.X., “A study of methods for negative relevance feedback,” Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 219-226.

[9] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” *Information Processing & Management*, vol. 24, 1988, pp. 513-523.

[10] S. Dominich, “Mathematical foundations of Information Retrieval,” Kluwer Academic Publisher, 2001.

[11] H. Chen and D.R. Karger, “Less is More Probabilistic Models for Retrieving Fewer Relevant Documents,” Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’06, New York, New York, USA: ACM Press, 2006, p. 429.

[12] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 335-336.

[13] B. Carterette and P. Chandar, “Probabilistic Models of Novel Document Rankings for Faceted Topic Retrieval,” Information Sciences, 2009, pp. 1287-1296.

[14] Y. Xu and H. Yin, “Novelty and topicality in interactive information retrieval,” Journal of the American Society for Information Science and Technology, vol. 59, Jan. 2008, pp. 201-215.

[15] E. Santos, Jr., H. Nguyen, J. T. Wilkinson, F. Yu, D. Li, K. J. Kim, R. Jacob, and O. Adam, “Capturing User Intent for Analytic Process,” Lecture Notes in Computer Science 5535: Proceedings of the User Modeling, Adaptation, and Personalization, 17th International Conference (UMAP 2009), 349-354, Trento, Italy, 2009.

[16] E. Santos, Jr., H. Nguyen, F. Yu, K. J. Kim, D. Li, J. T. Wilkinson, A. Olson, and R. Jacob, "Intent-Driven Insider Threat Detection in Intelligence Analyses," in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, 2008, pp. 345-349.

[17] E. Santos, Jr., H. Nguyen, F. Yu, D. Li, J. T. Wilkinson, “Impacts of Analysts' Cognitive Styles on the Analytic Process,” in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, 2010, pp. 601-610.

[18] D. D. Sleator and D. Temperley, "Parsing english with a link grammar," in Third International Workshop on Parsing Technologies, 1993.

[19] N. Chomsky, Syntactic Structures, 2nd ed. Mouton, Dec. 1957.