IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

Extrinsic Plagiarism Detection System for Semantic Replication in Medline

S. Sri Dharani[1], J. Ganesh[2], R. Ieshwarya[3], M. Sureka[4]

Students, Dept of Computer Science and Engg, Arasu Engineering College, Kumbakonam, TamilNadu, India[1,3,4]

Staff Member, Dept of Computer Science and Engg, Arasu Engineering College, Kumbakonam, TamilNadu, India[2]

Abstract: In a research area, plagiarism detection is more important to identifying duplicate documents in MEDLINE. In this paper ,we find the sentence based meaning. i.e., the given documents to match with the several documents for that if any sentences meanings are similar, we can find out easily. Each word has multiple meaning and multiple concepts (CUI) and also several alternative words to deal with the given documents. Information Retrieval based MEDLINE plagiarism detection has two approaches such as the candidate document selection and detailed analysis. The first attempt candidate document selection, identifying a set of candidate source from a document collection. In the second stage of detailed analysis, which make an complete comparison of the suspicious document with all candidates to identify similar sections. The Selected suspicious documented can also be check with the vocabulary mismatch by using Query Expansion. It's based on the UMLS Metathesarus and Word Sense Disambiguation. To identify the candidate document selection method by using Kullback-Leibler Distance.

Keyterms: Information Retrieval, Kullback-Leibler Distance, MEDLINE, Plagiarism Detection, UMLS Metathesarus, Word Sense Disambiguation.

I. INTRODUCTION

Academic area Plagiarism is major problem to copying of someone else information. It has two stages the first stage is Intentional plagiarism- the candidate should known the contents which are taken from some other authors. The second stage is Unintentional plagiarism- the candidate doesn't known which is taken from others documents. The plagiarism detection should improve the student knowledge and then they can learn about the several fields.

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a large collection of documents. The academic journals are Nursing, pharmacy, Healthcare etc. It can be analysis the each documents such as title, abstract, publisher date, etc. The candidate document selection to compare the each and every document and after comparison of given process it can be detail analysis the next stage. If anyone document is missing while comparing the candidate document we can't full fill the next stage the Query Expansion is based on the IR techniques.

More than 5500 biomedical journals are indexed in MEDLINE. New journals are not included automatically or immediately. Selection is based on the recommendations of a panel, the literature selection technical review committee based on scientific scope and quality of a journal[1]. The database contains information such as its name abbreviations and publisher about all journals included in Entrez including pubmed. The major roles of research areas are:

1. Content Analysis: Describing the contents of documents in a form suitable for computer processing;
2. Information Structures: Exploiting relationships between documents to improve the efficiency and effectiveness of retrieval strategies;
3. Evaluation: The measurement of the effectiveness of retrieval.

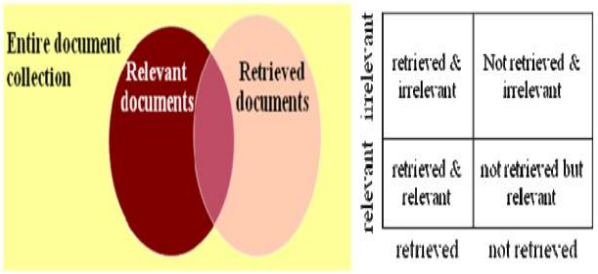

Fig 1. Document Collection based on IR

WSD task is a potential intermediate task for manyother NLP systems, including mono and multilingual Information Retrieval, Information Extraction, Machine Translation or Natural Language Understanding. WSD typically involves two main tasks.

i. Determining the different possible senses (or meanings) of each word.
ii. Tagging each word of a text with its appropriate sense with high accuracy and efficiency.

All methods build a representation of the examples to be tagged using some previous information.

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

The difference between them is the source of this information. The WSD community accepts a classification of these systems in two main general categories:

a) knowledge-based
b) corpus-based methods.

In Knowledge-based Method, mainly try to avoid the need of large amounts of training materials required in supervised methods[2][3]. Machine-Readable Dictionaries (MRDs) provide a ready-made source of information about word senses and knowledge about the world, which could be very useful for WSD and NLU.

MRDs contain inconsistencies and are created for human use, and not for machine exploitation. There is a lot of knowledge in a dictionary only really useful when performing a complete WSD process on the whole definitions.

Corpus-based Approach, these approaches are those that build a classification model from examples.These methods involve two phases: learning and classification. The learning phase consists of learning a sense classification model from the training examples. The classification process consists of the application of this model to new examples in order to assign the output senses. Most of the algorithms and techniques to build models from examples come from the Machine Learning area of AI.

One of the first and most important issues to take into account is the representation of the examples by means of features/attributes. That is, which information could and should be provided to the learning component from the examples. The representation of examples highly affects the accuracy of the systems. It seems to be as or more important than the learning method used by the system.

II. BACKGROUND

A. Knowledge Sources
Knowledge sources used for WSD are either lexical knowledge released to the public, or world knowledge learned from a training corpus.

1. Lexical Knowledge
In this section, the components of lexical knowledge are discussed. Lexical knowledge is usually released with a dictionary. It is the foundation of unsupervised WSD approaches.

2. Sense Frequency
It is the usage frequency of each sense of a word. Interestingly, the performance of the naïve WSD algorithm, which simply assigns the most frequently used sense to the target, is not very bad. Thus, it often serves as the benchmark for the evaluation of other WSD algorithms.

B. Learned World Knowledge
World knowledge is too complex or trivial to be verbalized completely. So it is a smart strategy to automatically acquire world knowledge from the context of training corpora on demand by machine learning techniques. The frequently used types of contextual features for learning are listed below.

1. Indicative Words
It surround the target and can serve as the indicator of target senses. In general, the closer to the target word, the more indicative to the sense. There are several ways, like fixed-size window, to extract candidate words.

2. Syntactic Features
It refer to sentence structure and sentence constituents. There are roughly two classes of syntactic features. One is the Boolean feature; for example, whether there is a syntactic object. The other is whether a specific word appears in the position of subject, direct object, indirect object, prepositional complement, etc. (Hasting, 1998; Fellbaum, 2001).

3. Domain-specific Knowledge
It is like selectional restrictions, is about the semantic restrictions on the use of each sense of the target word. However, domain-specific knowledge can only be acquired from training corpora, and can only be attached to WSD by empirical methods, rather than by symbolic reasoning. Hasting (1998) illustrates the application of this approach in the domain of terrorism.

4. Parallel Corpora
Parallel corpora is also called bilingual corpora, one serving as primary language, and the other working as a secondary language. Using some third-party software packages, we can align the major words (verb and noun) between two languages. Because the translation process implies that aligned pair words share the same sense or concept, we can use this information to sense the major words in the primary language.

There are no significant distinctions between lexical knowledge and learned world knowledge. If the latter is general enough, it can be released in the form of lexical knowledge for public use. Usually, unsupervised approaches use lexical knowledge only, while supervised approaches employ learned world knowledge for WSD. Examining the literature, however, we found the trend of combination of lexical knowledge and learned world knowledge in recently developed WSD models.

III. PROPOSED APPROACH

This section presents the IR-based approach to the identification of candidate source documents followed by a description of how it can be extended by query expansion using resources from the medical domain.

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

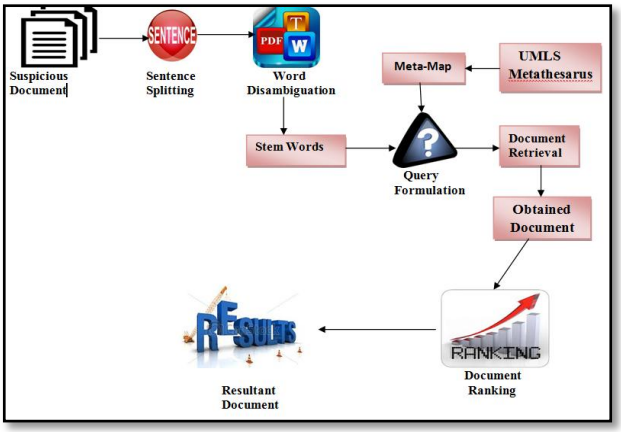

Fig 2. The proposed architecture for plagiarism detection

A. IR-Based Approach

The process of retrieving candidate source documents using the proposed IR-based approach.The source collection is indexed with an IR system. In the IR-based framework, candidate retrieval process can be divided into four main steps pre-processing,(2) query formulation,(3) retrieval and (4) results merging. These steps are described as follows:

1. Pre-Processing:

Each suspicious document is split into sentence using NLTK. The terms in each sentence are converted to lower case.stopword and punctuation marks are removed.

2. Query Formulation:

Sentences from the suspicious document are used to form multiple queries.The length of a query can vary from a single sentence to all sentences appearing in a document as reused text can be sourced from one or more documents and vary from a single sentence to an entire document . A long query is likely to perform well in situations when large portions of text are reused for plagiarism; on the other hand small portions of plagiarized text are likely to be effectively detected by a short query. Therefore, the choice of query length is important in obtaining effective results.

3. Retrieval:

Terms are weighted using the tf.idf weighting sceme and then text forming the query is used to retrieve similar documents from the index.

4. Result Merging:

The top N documents returned against multiple queries are merged to generate a final ranked list of source documents. A standard data fusion approach, CombSUM, is used to generate the final ranked list of documents by combining the similarity scores of source documents retrieved against multiple queries.

In CombSUM the final similarity score, Sfinalscore, is obtained by adding the similarity scores of source documents obtained against each query q:

$$S_{finalscore} = \sum_{q=1}^{Nq} Sq(d) \qquad (1)$$

Where,

Nq is the total number of queries to be combined.

Sq(d) is the similarity score of a source document d for a query q.

The top K documents in the ranked list generated by the CombSUM method are marked as potential candidate source documents

B. Query Expansion

The Unified Medical Language System4 (UMLS), a set of tools and resources to assist with the development of biomedical text processing systems, is used to carry out query expansion. Our approach uses two main UMLS resources (the Metathesaurus and MetaMap) which are now described,followed by an explanation of how they are used for queryexpansion.

1. UMLS Metathesaurus:

The UMLS Metathesaurus is a large database of more than 100 multi-lingual controlled source vocabularies and classifications, which contains information about concepts (related to biomedical and health), concept names and relationships between concepts[4][5]. The basic units of the Metathesaurus are concepts, whereby the same concept can be referred to using different terms.

One of the main goals of Metathesaurus is to group all the equivalent terms (synonyms) from different source vocabularies into a single concept. Thus, a concept is a collection of synonymous terms. Each concept in Metathesaurus is assigned a unique identifier called a CUI (Concept Unique Identifier).

2. MetaMap using Full Read source vocabulary

As well as diagnoses, the full Read dictionary contains terms for temporality, laterality, body parts etc. which can match fragments of text in isolation but may not convey clinically useful information (for example, the Read term ‘Disease’ could match any mention of the word “disease” in the text). Therefore we restricted the output to Read terms with the following semantic types and other Read terms extracted from the same phrase (which might give additional contextual information).

Example:

- Acquired Abnormality
- Acquired Abnormality, Disease or Syndrome
- Anatomical Abnormality
- Congenital Abnormality
- Disease or Syndrome
- Environmental Effect of Humans, Hazardous or Poisonous substance
- Finding
- Injury or Poisoning
- Laboratory or Test Result
- Mental or Behavioral Dysfunction

DOI 10.17148/IJIREEICE.2016.41108

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

- Mental Process
- Neoplastic Process
- Organ or Tissue Function
- Pathologic Function
- Phenomenon or Process
- Sign or Symptom

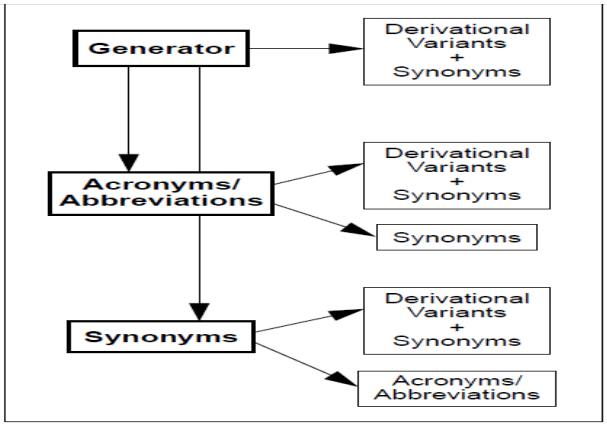

Fig 3. Find out the changes in Abbreviations

C. Candidate Retrieval

The candidate set of all Metathesaurus strings containing at least one of the variants is retrieved[6].This retrieval is controlled by various options including stop_large_n which precludes searching for candidates containing either singlecharacter variants with more than 2,000 occurrences in the Metathesaurus and two-character variants with more than 1,000 occurrences. In addition candidate retrieval is made more efficient through the use of special, small indexes whenever possible.

D. Candidate Evaluation

Each Metathesaurus candidate is evaluated against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics:

1. centrality (involvement of the head),
2. variation (an average of inverse distance scores),
3. coverage and
4. cohesiveness.

The latter two components measure how much of a candidate matches the text and in how many pieces. The candidates are then ordered according to mapping strength.

E. Impact Of Different Matching Strategies To Disambiguation Quality

To test the effectiveness of different matching strategies, we performed some additional experiments. The disambiguation results by each individual document with the following 5 matching strategies:

1. Dependency matching only.
2. Dependency and backward matching.
3. Dependency and synonym backward matching.
4. Dependency and synonym dependency matching.
5. Dependency, backward, synonym backward, and synonym dependency matching.

As expected combination of more matching strategies results in higher disambiguation quality[7][8]. By analyzing the scoring details, we verified that backward matching is especially useful to disambiguate adjectives and adverbs. Adjectives and adverbs are often dependent words, so dependency matching itself rarely finds any matched words. Since synonyms are semantically equivalent, it is reasonable that synonym matching can also improve disambiguation performance.

E. The Knowledge Discovery Process

Data mining is one of the tasks in the process of knowledge discovery from the database. The data stored in the database is used to discover the patterns of data, which then interpreted by applying the domain knowledge. Following figure shows the process of Knowledge Discovery from Database.

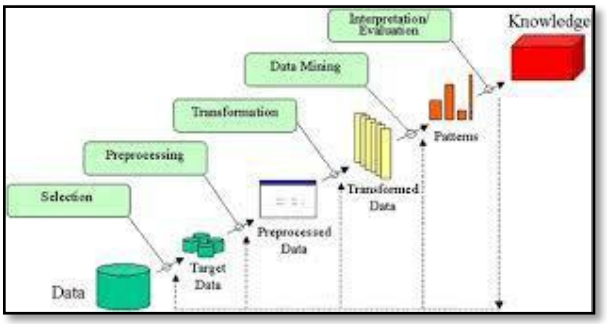

Figure 4: The Knowledge Discovery Process

1. Stemming Algorithm

A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form, for example,

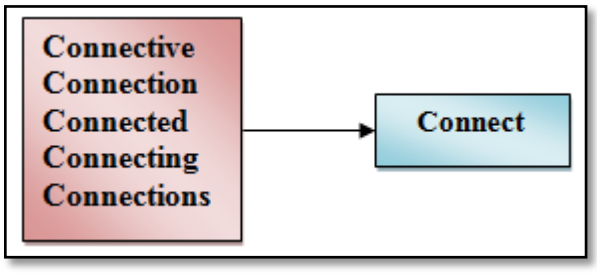

Figure 5: Stem Word

It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes. In addition,

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

stemming algorithms - at least the ones presented here - are applicable to the written, not the spoken, form of the language.

IV. EXPERIMENTAL SETUP

This section describes the dataset used for evaluation and how the approach was implemented and the evaluation measure used to evaluate the various query expansion methods.

A. Evaluation Dataset

Evaluation is carried out using an existing source of potentially plagiarised publications from Medline. For these experiments, the source collection is fromed from 19,569,568 citations from the 2011MEDLINE/PubMed Baseline Repository. The collection of suspicious documents contains 260 citations from the Deja vu database that have been manually examined and verified as duplicates. These citation pairs are selected because they do not have a common author, making them potential cases of plagiarism.

B. Implementation

Lucene8, a popular and freely available IR system, is used for the experiment. The source collection is indexed. Documents are pre-processed by converting the text into lower case and removing all non-alphanumeric characters. Stopwords9 are removed and stemming is carried out using the Porter Stemmer.Terms are weighted using the tf.idf weighting scheme.Lucene computes the similarity score between query and document vectors.

Lucene computes the similarity score between query and document vectors using the cosine similarity measure:

$$sim(d,q) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}| \times |\vec{d}|} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2 \times \sum_{i=1}^{n}(d_i)^2}} \quad \textbf{(2)}$$

where $|\vec{q}|$ and $|\vec{d}|$ represent the lengths of the query and document vectors respectively.

C. Evaluation Measure

The goal of the candidate document retrieval task is to identify all the source document(s) for each suspicious document while returning as few non-source documents as possible. It is important for all source documents to be included in the top ranked documents returned by the system since otherwise they will not be identified during later stages of processing[9]. Consequently, recall is more important than precision for this problem. Recall for the top K document, averaged across queries is used as the evaluation measure for these experiments[10]. For a single query the Recall at K (R@K) is 1 if the source document appears in the top K documents retrieved by the query, and 0 otherwise. For a set of N queries, the averaged recall at K score is calculated as:

$$R@K_{avg} = \frac{1}{|N|}\sum_{i=1}^{N} R@K_i \quad \textbf{(3)}$$

where R@Ki is the recall at K score for query i.

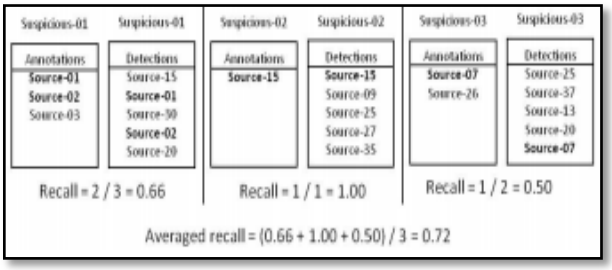

Fig 6. Example showing calculation of averaged recall score

V. RESULT ANALYSIS

Our proposed IR-based approach for retrieving candidate documents performs well in identifying real cases of plagiarism. Performance further improves when query expansion is applied. Table1 shows the results of the experiments for the top 1, 5, 10, 15 and 20 candidate source documents. As expected, retrieval performance increases as the number of retrieved documents increases.

TABLE I PERFORMANCE FOR THE MEDLINE CORPUS

	Avg. Recall for top K documents				
Approach	1	5	10	15	20
Kullback-Leibler	0.7596	0.8154	0.8442	0.8558	0.8596
No Query Expansion	0.8769	0.9173	0.9250	0.9288	0.9288
WSD	0.9077	0.9519	0.9558	0.9558	0.9596
Without-WSD	0.9035	0.9519	0.9519	0.9558	0.9558
WSD Phrase	**0.9219**	**0.9595**	0.9595	**0.9652**	0.9652
Without-WSD Phrase	0.9115	0.9558	**0.9596**	0.9634	**0.9673**

Performance is compared against the the Kullback-Leibler Distance method[11][12]. This approach is based in pairwise comparison of documents which would be computationally expensive for the source collection of over 19 million citations used by the IR-based approach. Consequently a randomly selected subset of 3 million citations, which include the sources for the 260 plagiarised citations, is used as source collection for experiments with the Kullback-Leibler Distance approach.

TABLE II QUERY PERFORMANCE

		No. of Queries (%) effecting Rank		
Corpus	Approach	Higher	Lower	Same
MEDLINE	WSD	14(5.38)	2(0.77)	234(90.00)
	Without-WSD	17(6.54)	5(1.92)	230(88.46)
	WSD Phrase	13(5.00)	4(1.54)	234(90.00)
	Without-WSD Phrase	15(5.77)	4(1.54)	233(89.62)

DOI 10.17148/IJIREEICE.2016.41108

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

Note that an implication of this decision is that the Kullback-Leibler Distance approach has the advantage of a significantly smaller search space from which to identify source documents.

The IR-based approach proposed here achieves higher results than the Kullback-Leibler Distance approach. Highest recall achieved by this method is 0.8596 for top 20 candidate documents, although it is expected that performance will drop when the entire MEDLINE database is used. The proposed approach (without query expansion) achieves a recall of 0.8769 for K = 1, which is still higher than the maximum recall obtained using the Kullback-Leibler Distance method. This high recall score indicates the strength of the proposed method in detecting potential real cases of plagiarism from large reference collections. As expected, retrieval performance improves when query expansion is applied. Improvement in performance is statistically significant for all query expansion approaches (Wilcoxon signed-rank test, $p < 0.05$)[13].

VI. CONCLUSIONS

In this paper, we introduced a novel framework for query formulation. This framework synthesizes, in a principled and effective manner, arbitrary concept matches, concept weighting and query expansion. Our query formulation approach leverages external sources of information such as web n-gram counts, anchor and heading text extracted from a large web corpus, and articles and titles from Wikipedia for weighting the explicit query concepts as well as selecting relevant and diverse set of weighted expansion terms. You can search for some specific topic, some job seeking activities, and project related searches and of course business and finance oriented statistical search as well.

REFERENCES

[1] Rao Muhammad Adeel Nawab,Mark Stevenson,Paul Clough.An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE.IEEE/ACM Transaction on Computational Biology and Bioinformatics.(volume pp,Issue:99),March 2016.

[2] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4):357–389, October 2002.

[3] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In Proc. of SIGIR, pages 571–578, 2010.

[4] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In Proc. of SIGIR, pages 491–498, 2008.

[5] M. Bendersky, D. Fisher, and W. B. Croft. UMass at REC 2010 Web Track: Term Dependence, Spam Filtering and Quality Bias. In Proc. of TREC-10, 2011.

[6] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In Proc. Of WSDM, pages 31–40, 2010.

[7] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In Proc. of SIGIR, 2011.

[8] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In Proc. Of SIGIR, pages 243–250, 2008.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In Proc. Of SIGIR, pages 659–666, 2008.

[10] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In Proc. of TREC-09, 2010.

[11] W. B. Croft, M. Bendersky, H. Li, and G. Xu. Query representation and understanding workshop report. SIGIR Forum, December 2010.

[12] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In Proc. of SIGIR, pages 154–161, 2006.

[13] J. Guo, G. Xu, H. Li, and X. Cheng. A unified and discriminative model for query refinement. In Proc. of SIGIR, pages 379–386, New York, NY, USA, 2008.