Query Wikification: Mining Structured Queries From Unstructured Information Needs using Wikipedia-based Semantic Analysis

Amir Hossein Jadidinejad, Fariborz Mahmoudi
Islamic Azad University of Qazvin
amir@jadidi.info, mahmoudi@itrc.ac.ir

Abstract

Combining the language model and inference network, as implemented in the Indri search engine, is efficient and verified approach. In this retrieval model, the user's information need is exhibited as Indri's Structural Query Language. Although the SQL allows expert users to richly represent its information needs but unfortunately, the complicacy of SQLs make them unpopular in the WEB for ordinary ones. Automatically detecting the concepts in a user's information need and generate a richly structured equivalent query is a good solution. It needs a concept repository and a way to extracting appropriate concepts from the user's information need. We utilize Wikipedia as a great, multilingual, free-content encyclopedia for our knowledge base and also some state of the art algorithms for extracting Wikipedia's concepts from the user's information need. This process is called "Query Wikification". Mining Wikipedia concept repository help us to propose a solution that supports usability in multilingual environments, cross-language retrievals, scalability and covering erratum, various equivalents and synonyms of a concept. Experimental results verify that our automatic structured query construction is an efficient and scalable method that has a very good potential to apply on the WEB. Our experiments over TEL corpus in CLEF2009 achieves **+23%** improvement in Mean Average Precision and retrieves more than **600** relevant documents against the Indri baselines.

In Persian track, we evaluated a simple stemmer so-called "Perstem", a stemmer and light morphological analyzer for Persian language. Our experimental results show that using this stemmer in indexing and retrieval phase can significantly improve both precision (+**91%**) and recall (+**43%**).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Indri Structure Query Language, Wikipedia knowledge-base, Meta-Language Indexing, Query Wikification, WM-Wikifier, Perstem.

1 Introduction and Motivation

Representing user's information need is a fundamental part in an information retrieval system. Most systems get a list of keywords for each information need. For example, if a user is interested in "colour therapy" and the therapeutic use of colour they might formulate the natural language query ***"colour therapy"***. It's not only a hard task for ordinary users to represent its information need as a set of keywords but also clear that a lot of semantics is lost by transcribing the information need into a set of keywords. Such a query may retrieve some documents about "color" or "therapy" that completely irrelevant. Also user's knowledge about the query is neglected when encoding it as a list of keywords. For example, maybe our user knows that "color" and "colour" are synonymous!

Structured Queries can represent user's information needs accurately. A Structured Query Language (SQL) allows terms weighting, the use of proximity information among terms, field restricting and various ways of combining concepts. Since structured queries can be more expressive than keywords, it's verified that retrieval models that can evaluate structured queries [8] such as Indri [12] and InQuery [3] have more potential to retrieve more accurate results.

Although the structured queries and related models got a very good results in different experiments [12, 8] but they suffer from a drawback that made them unusable in the WEB. *Having knowledge about related concepts in the query is necessary to constructing structured queries. Even we presume that the user has a good knowledge about its information need, learning the complicated Structured Query Language for WEB users is not favorable.* Understanding the user's information need and generating a richly structured query is a great solution. It needs a huge concept repository that covers all query concepts and a way to extracting appropriate concepts from the user's information need. Wikipedia is a multilingual, web-based, free-content encyclopedia that cover most important concepts in the world. We call the process of extracting a list of Wikipedia concepts from a natural language information need as "Query Wikification". Mining Wikipedia and some state of the art Wikification algorithms [9, 10] are used to generate a richly, efficient structured query. The contributions of this paper are the following:

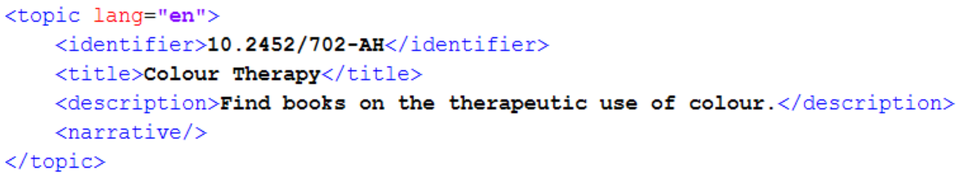
- Proposing a new method for converting a simple natural language information need into a well-formed, rich, efficient, structured query. This process is done with the aid of state of the art algorithms in both "Wikification" [9, 10] and "Structural Retrieval Models" [8]. It can replace keyword-based search engines on the WEB with the powerful structured queries and related models.

- Usability in multilingual environments and cross-language retrieval. The proposed model make a meta-language search engine from Indri [12] that can efficiently apply on multilingual environments such as the WEB. Our experiments in CLEF2009 campaign is a good evidence for this feature.

- Scalability. The proposed approach is base on Indri search engine[1], a scalable language modeling search engine that supports structured queries. Also some new projects such as Galago[2] that supports Indri Structured Query Language in a distributed computation framework make it more and more scalable and suitable for the WEB.

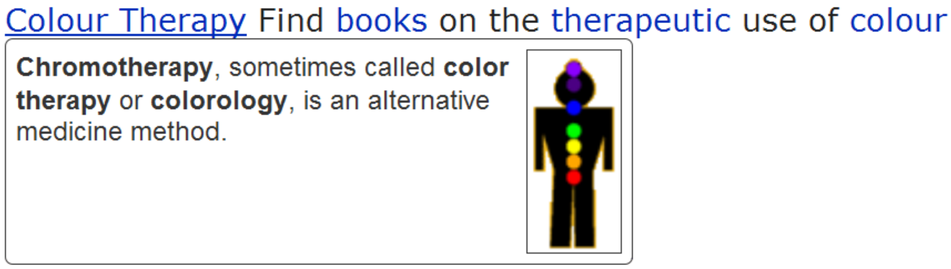
- Our model can extract a vocabulary for each query's concept by mining Wikipedia. It contains erratum, stemmed equivalent and synonyms of the concept. All of them are embedded in the structured query. Also, this vocabulary can work as a semi-stemming algorithm and very helpful in multilingual environments or complicated languages such as Persian language that have a hard morphology (Sec. 4.2.2). This feature is WEB suitable too!

[1] http://lemurproject.org/indri/
[2] http://www.galagosearch.org/



```
<topic lang="en">
    <identifier>10.2452/702-AH</identifier>
    <title>Colour Therapy</title>
    <description>Find books on the therapeutic use of colour.</description>
    <narrative/>
</topic>
```

Figure 1: A sample user's information need (NO. 10.2452/702-AH)



Colour Therapy Find books on the therapeutic use of colour

Chromotherapy, sometimes called **color therapy** or **colorology**, is an alternative medicine method.

Figure 2: A sample user's information need after Query Wikification (NO. 10.2452/702-AH)

2 Query Wikification

The process of automatically recognizing the topics mentioned in unstructured text and linking them to the appropriate Wikipedia articles is known as wikification [9]. The user's information need is a short and informative text. So we can apply Wikification on user's information needs in order to map unstructured query into a weighted list of concepts in Wikipedia. We call this process as "Query Wikification". To our knowledge, there isn't any relevant publication in this research area.

Two Wikification method have been proposed by now. The first is Wikify! [9] and the second is WM-Wikifier [10]. WM-Wikifier is a distinguish approach that uses Wikipedia articles not only as a source of information to point to, but also as training data for how best to create links. We utilize this algorithm for "Query Wikification". More details can be found in [10].

For example, take a look at Figure 1. It's a sample user's information need in CLEF 2009. The result of Query Wikification is shown in Figure 2. As you see, the important topics are extracted and the original query is annotated using Wikipedia concepts. We use Wikipedia-Miner[3] toolkit [14] in our experiments.

3 Structured Query Construction

If we can map an unstructured user's information need to a weighted list of Wikipedia concepts, what can we do with these concepts?!, It can help us to move from unstructured, limited and noisy text to structured, well-known and accurate concepts. It's a break through step in Information Retrieval. The Wikification algorithms simply do that!

In our experiments, we utilize the WM-Wikifier [10] algorithm in order to extract a weighted list of Wikipedia concepts and mine translation and synonyms of these concepts from Wikipedia knowledge-base to construct an equivalent structure query. For example, take a look at Figure 1. It's a sample topic in CLEF 2009. In this topic the user is looking for all relevant information bout colour therapy and therapeutic use of colours. The following is the Indri [12, 8] equivalent structure query after removing redundant and stop words:

```
#combine(colour therapy therapeutic)
```

[3] http://wikipedia-miner.sourceforge.net/

Title	Distribution	Description
dc:title	80%	This is record's title. All records contains this field and it ia a valuable field.
dcterms:alternative	little	In some records, this field contains relevant information.
dc:subject	210%	Manually assigned subject heading.
dc:abstract	little	Record's abstract.
dc:description	42%	Record's description. Mostly contains copyrights and related stuffs.
dc:contributor	little	Record's contributor.

Table 1: Valuable fields in preprocessing step.

The following structure query is generated by our approach[4]. It contains some professional expressions ("chromotherapy") and all translations and synonyms of each concept:

```
#combine(colour therapy therapeutic
  #syn(chromotherapy farbtherapi colourology #1(color therapy))
  #syn(color couleur farb colour colors colours couleur)
  #syn(therapi thrap therapi treatment therapie therapy))
```

There are various approach in constructing equivalent structure query. In the next section, we describe our experiments.

4 Experiments

4.1 TEL@CLEF2009

4.1.1 Meta-Language Field Index Construction

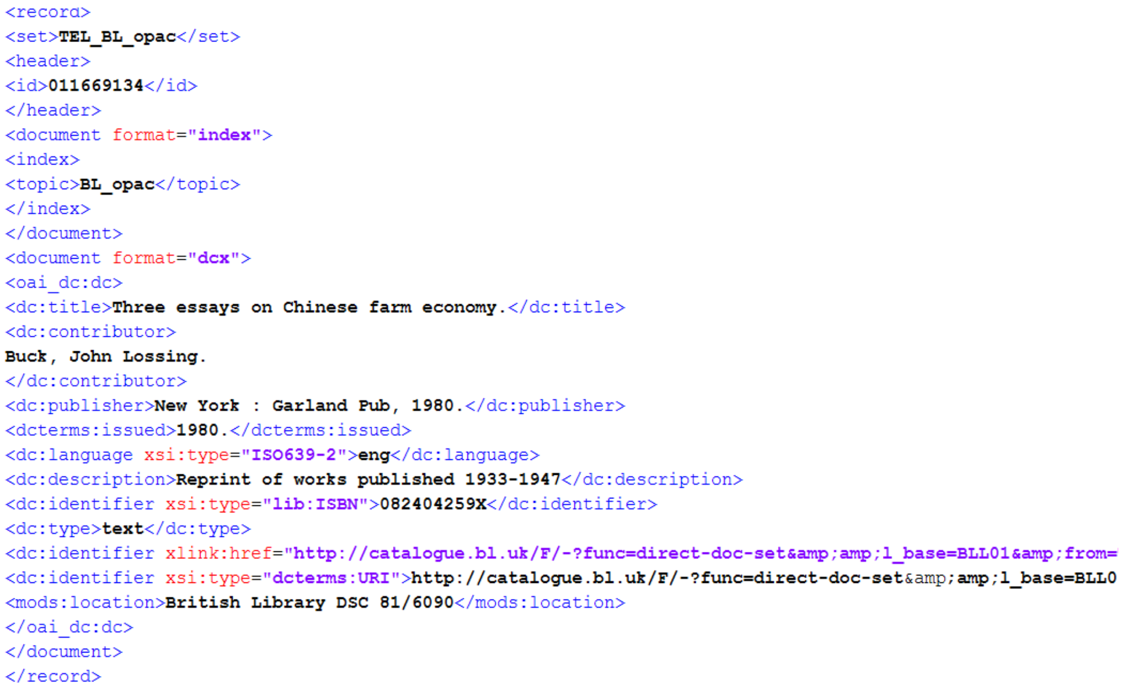
TEL is an inherently multilingual corpus. It contains not only records in different languages but also some records maybe have multilingual fields. Detecting record's language is a fundamental task to apply stemming and stop word removal. On the other hand, detecting different languages in each record is not only a hard work but also lead to poor results. Previous experiments utilize different language identification approaches to detect each field's language and then apply appropriate stemmer and stop words [1]. We use a meta-language index in our experiments. Instead of distinguishing different languages, all fields are indexed without stemming and stop word removal. In this approach, all valuable contents are indexed together without any concern about underlying language. It is clear that such indexing strategy is not appropriate in general but our experiments have shown that it is an appropriate indexing strategy in tandem with Query Wikification and Indri Structured Query Language.

In the preprocessing step, we delete all noisy and invaluable fields from TEL corpus. After analyzing TEL's records, we extract a list of fields that contains important information. Table 1 shows the valuable fields in preprocessing step. For example, see Figure 3, it is a sample record in TEL corpus. Figure 4 is an equivalent record after preprocessing. As you see, we skip all invaluable fields and store remaining one in TREC format. Also we don't apply any stemming or stop word removal in the indexing phase. Instead apply stop word removal in retrieval phase using a list of stop words provided by UNINE[5].

We utilize Indri [12, 8] Field Index for indexing because it not only construct a powerful field index but also support index's fields in its query language. All valuable fields (Table 1) is configured

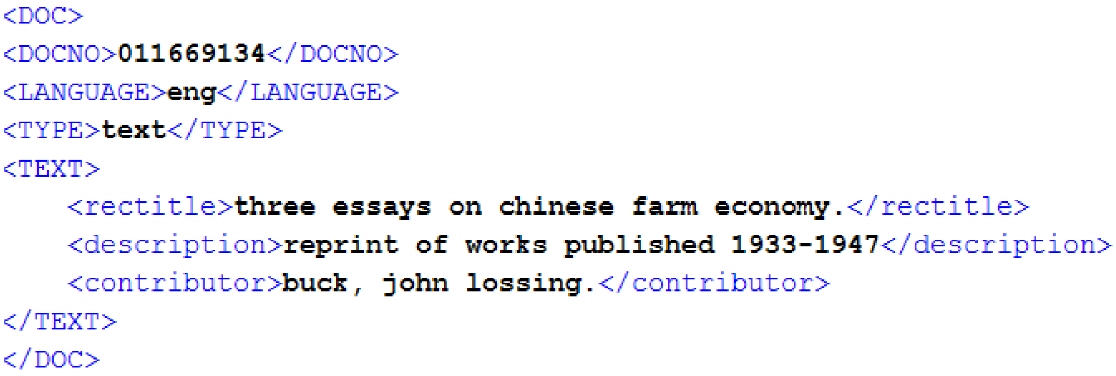
[4] The generation procedure is discussed in Sec 4.

[5] http://members.unine.ch/jacques.savoy/clef/englishST.txt



```
<record>
<set>TEL_BL_opac</set>
<header>
<id>011669134</id>
</header>
<document format="index">
<index>
<topic>BL_opac</topic>
</index>
</document>
<document format="dcx">
<oai_dc:dc>
<dc:title>Three essays on Chinese farm economy.</dc:title>
<dc:contributor>
Buck, John Lossing.
</dc:contributor>
<dc:publisher>New York : Garland Pub, 1980.</dc:publisher>
<dcterms:issued>1980.</dcterms:issued>
<dc:language xsi:type="ISO639-2">eng</dc:language>
<dc:description>Reprint of works published 1933-1947</dc:description>
<dc:identifier xsi:type="lib:ISBN">082404259X</dc:identifier>
<dc:type>text</dc:type>
<dc:identifier xlink:href="http://catalogue.bl.uk/F/-?func=direct-doc-set&amp;amp;l_base=BLL01&amp;from=
<dc:identifier xsi:type="dcterms:URI">http://catalogue.bl.uk/F/-?func=direct-doc-set&amp;amp;l_base=BLL0
<mods:location>British Library DSC 81/6090</mods:location>
</oai_dc:dc>
</document>
</record>
```

Figure 3: A sample record in TEL corpus.



```
<DOC>
<DOCNO>011669134</DOCNO>
<LANGUAGE>eng</LANGUAGE>
<TYPE>text</TYPE>
<TEXT>
    <rectitle>three essays on chinese farm economy.</rectitle>
    <description>reprint of works published 1933-1947</description>
    <contributor>buck, john lossing.</contributor>
</TEXT>
</DOC>
```

Figure 4: A sample record after preprocessing.

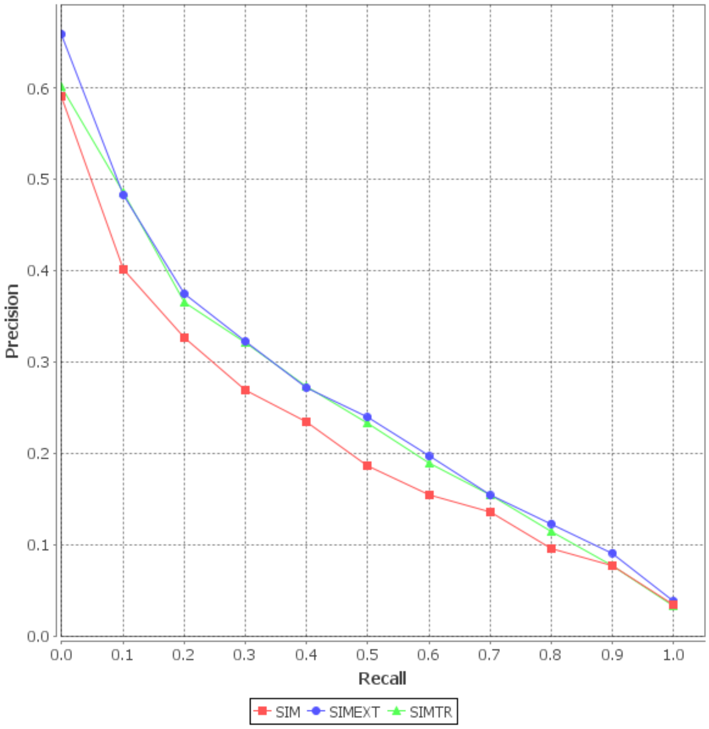


Figure 5: Precision/Recall Graph: Performance evaluation for all queries

as a backward field index. Finally, the indexing is done by “indribuildindex” application in Lemur toolkit [11].

4.1.2 Indri Baseline

To compare our results, we apply Indri retrieval model [12, 8] on the *title* and *description* of each topic. The query model is as follow:

```
#combine( <title> <description> )
```

Before passing topics to Indri retrieval engine, all common and redundant words are removed. For example, for the query that is shown in Figure 1, after removing common and redundant words:

```
#combine(colour therapy therapeutic)
```

This run is addressed as “SIM” in the our experiments. Table 3 and Figures 5 and 6 compare this baseline with proposed approaches.

4.1.3 Concept Translation

Wikipedia contains articles in more than 250 natural languages. Each article link to equivalent one in other languages. After extracting concepts from unstructured user’s information need, we can utilize the translation links in Wikipedia in order to translate each concepts. The following model is applied:

```
#combine( <title> <description> #syn(#1(EN) #1(FR) #1(GE)) )
```

For example for previous sample query:

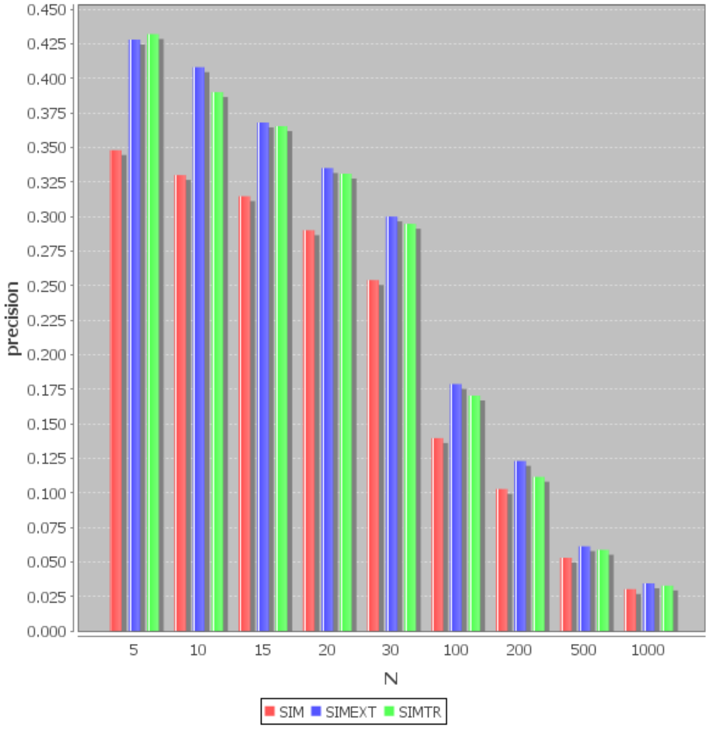


Figure 6: Precision@N Graph: Performance evaluation for all queries

```
#combine(colour therapy therapeutic
                #syn(chromotherapy farbtherapi)
                #syn(color couleur farb)
                #syn(therapi thrap therapi))
```

This run is addressed as "SIMTR" in our experiments. Table 3 and Figures 5 and 6 compare this run with other approaches and the baseline. Also take a look at Table 2, it compares the proposed approaches and baseline for the previous example ("colour therapy"). Evaluation results show that translating concepts using Wikipedia significantly improve both precision (+18%) and recall (+8%). For the example query (Table 2), Mean Average Precision is improved (+62%) and also 1 (+4%) more relevant document is retrieved.

4.1.4 Concept Translation and Synonyms Extraction

Most retrieval systems are a simple pattern matcher. So co-occur terms play an important role in ranking algorithm. So we eager to know more and more synonyms and relevant concepts for each concept. If we have an article in Wikipedia, we can mine all other articles to find a list of synonyms for this article. There are two distinct ways: redirect pages[6] and anchors. We prefer anchor titles since we can rank the vocabulary for each concept while ranking is not possible for redirect pages[7]. This can be done by anchor texts. All anchors for one articles are synonym. This assumption construct the following structure query:

```
#combine( <title> <description>
  #syn(#1(EN) #1(FR) #1(GE) <Anchors List>))
```

For example the previous sample query is defined as:

[6] redirects are standalone pages in Wikipedia that just have a title that refer to an article. For covering various equivalents, misspelling, and...

[7] We can rank redirect pages by query logs in Wikipedia.

RUN	Relevant-Retrieved	MAP	NDCG	R-PREC
SIM	25/29	0.4964	0.7793	0.4138
SIMTR	26/29	0.8043	0.9079	0.8276
SIMEXT	27/29	0.8230	0.9239	0.8276

Table 2: Performance evaluation for 10.2452/702AH query.

RUN	Relevant-Retrieved	MAP	NDCG	R-PREC
SIM	1518/2527	0.2013	0.4635	0.2350
SIMTR	1645/2527	0.2390	0.5132	0.2688
SIMEXT	1724/2527	0.2462	0.5306	0.2794

Table 3: Performance evaluation for all queries in monolingual TEL track.

```
#combine(colour therapy therapeutic
  #syn(chromotherapy farbtherapi colourology #1(color therapy))
  #syn(color couleur farb colour colors colours couleur)
  #syn(therapi thrap therapi treatment therapie therapy))
```

This run is addressed as "SIMEXT" in our experiments. Table 3 and Figures 5 and 6 compare this run with other approaches and the baseline. Also take a look at Table 2, it compares the proposed approaches and baseline for the previous example ("colour therapy"). Evaluation results show that translating concepts in tandem with synonyms and various equivalent extraction using Wikipedia significantly improve both precision (+22%) and recall (+13%). For the example query (Table 2), Mean Average Precision is improved (+66%) and also 2 (+8%) more relevant document is retrieved. Also our experimental results over TEL corpus show that SIMEXT is a better solution than SIMTR in both precision and recall.

4.2 Persian@CLEF

Persian is an Indo-European language spoken in Iran, Afghanistan and Tajikistan. It is also known as Farsi [1, 6]. In this section we summarize our experiments in the Persian track of CLEF2009.

4.2.1 Bilingual

Bilingual retrieval in Persian track is done with a same approach as discussed in Sec. 4.1. Unlike TEL experiments, we have a very poor results, due to little coverage of Farsi language of Wikipedia[8]. For example most topics is extracted from the query but since there isn't an equivalent article in Farsi language of Wikipedia, we can't translate it. Table 4 shows our different runs.

[8] http://fa.wikipedia.org

RUN	Relevant-Retrieved	MAP	NDCG	R-PREC
IAUPEREN1	650/4330	0.0195	0.0975	0.0433
IAUPEREN2	659/4330	0.0202	0.0984	0.0427
IAUPEREN3	773/4330	0.0277	0.1223	0.0477

Table 4: Performance evaluation for all queries in bilingual Persian track.

RUN	Relevant-Retrieved	MAP	NDCG	R-PREC	Desc
IAUPERFA1	3528/4464	0.3459	0.6674	0.3750	Stemmed, PRF(5,10)
IAUPERFA2	2403/4464	0.0202	0.4268	0.2083	No stemming, PRF(5,10)
IAUPERFA3	3820/4464	0.3762	0.7089	0.4033	Stemmed
IAUPERFA4	2670/4464	0.1964	0.4649	0.2345	Indri Baseline

Table 5: Performance evaluation for all queries in monolingual Persian track.

4.2.2 Monolingual

Perstem[9] is a stemmer and light morphological analyzer for Persian by *Jon Dehdari*[10]. It is written in Perl and uses regular expression substitutions to separate inflectional morphemes and remove affixes. The stemmer currently has 76 substitution rules, which replace one pattern of text with another [4]. It has a very good performance and accuracy for stemming and morphological analyzing of Persian texts. On a sample dataset, Perstem correctly and efficiently analyzed 97% of the words [4].

Inconsistent stemming results have been reported in CLEF2008 [2, 7]. So we decided to evaluate it in our CLEF 2009 experiments. Unlike [13], our evaluation is based on overall performance (precision/recall) with Hamshahri corpus and benchmark queries in CLEF 2009. On the other hand, we investigate the application of Perstem in Persian retrieval in a large news corpus. Table 5 shows our official runs[11]. Experimental results show that stemming algorithm significantly improved both precision (**+91%**) and recall (**+43%**).

5 Conclusion and Future Works

In this paper we propose an efficient approach for extracting relevant concepts and a vocabulary of synonyms, translations, various equivalents and. . . that all of them are embedded in a structured query. We leverage Wikipedia as our knowledge base and Indri as Structured Query Language and model. Query modification techniques such as query expansion suffer from a problem so-called "Query Drift". It means that although by modifying a query we can get more relevant documents but it maybe hurt the precision. Our experiments over TEL corpus show that this method is an efficient and robust approach that significantly improves both precision and recall. We believe that our method is a good potential to apply on the WEB. For example, take a look at the following query[12]:

```
Title: Modern Persian Language,
Desc: Retrieve publications providing instructions on
learning or teaching modern/contemporary Persian.
```

Take a look at the generated structured query by *SIMEXT*:

```
#weight(0.3 #combine(modern teaching instructions persian contemporary learning language)
0.7 #syn(farsi #1(persian languages) #1(farsi salis language) #1(modern perisan) persian
#1(modern persian language) #1(parsi language) #1(farsi language) #1(modern persian)
#1(persian language) #1(persische sprache) ))
```

"Farsi" or "Parsi" are informal equivalents of "Modern Persian Language" that it can't nowise understand from the original query. Using these informal equivalent on the WEB is very important evidence. For another example, take a look at the following structured query for Figure 1:

[9] http://sourceforge.net/projects/perstem/

[10] http://www.ling.ohio-state.edu/~jonsafari/

[11] DOI: 10.2415/AH-PERSIAN-MONO-FA-CLEF2009.QAZVINIAU.IAUPERFA¡X¿

[12] 10.2452/733AH

```
#combine(colour therapy therapeutic
  #syn(chromotherapy farbtherapi colourology #1(color therapy))
  #syn(color couleur farb colour colors colours couleur)
  #syn(therapi thrap therapi treatment therapie therapy))
```

As you see, without applying a complicated stemmer in our multilingual environment (TEL corpus), our extracted vocabulary from anchor titles can cover most of them efficiently. For example, in the structured query, “color” and “colour” are synonyms. It’s a very good potential in highly multilingual environments such as the WEB. Evaluation comparison for each query is shown in Table 6.

6 Acknowledgements

We would like to thank Donald Metzler[13], one of the main developers of Indri Structured Query Language, for his ideas and advice, and Lemur community[14] for supporting and sharing an excellent resource. Also, we would like to thank Jon Dehdari for sharing Perstem, and DBRG[15] for Hamshahri corpus. Finally, we must of course acknowledge the tireless efforts of the Wikipedia community that make a valuable knowledge base during years. We are also debated to the CLEF organizers too.

[13] http://research.yahoo.com/Don_Metzler

[14] http://sourceforge.net/forum/?group_id=161383

[15] http://ece.ut.ac.ir/DBRG/

A CLEF2009 Query Details

Table 6: Query Details

ID	Title	R	Wikipedia Concepts		SIM		SIMTR		SIMEXT	
			Title	W	R	MAP	R	MAP	R	MAP
01	Arctic Animals	21	Fauna Arctic Arctic Ocean Species Animal	0.72 0.69 0.58 0.51 0.40	21	0.05	21	0.07	21	0.24
02	Colour Therapy	29	Chromotherapy Color Therapy	0.90 0.12 0.11	25	0.49	26	0.80	27	0.82
03	Chess for Beginners	36	-	-	36	0.44	36	0.44	36	0.44
04	Social Benefits of Sport	54	Social welfare provision Sport Activity Benefit Social Sporting Clube de Portugal	0.42 0.32 0.21 0.20 0.18 0.17	39	0.02	40	0.06	9	0.00
05	Volcanoes and Volcanism	84	Volcano	0.29	82	0.28	84	0.26	84	0.36
06	Caste System in India	107	Caste system in India India Caste Indian independence movement	0.96 0.95 0.78 0.23	86	0.54	86	0.53	105	0.59
07	Fantasy Role-playing Games	29	Role-playing game Fantasy List of fantasy subgenres Video game Roleplaying	0.85 0.68 0.58 0.37 0.29	29	0.48	29	0.46	29	0.51
08	Wedding Planning	67	Wedding ceremony participants Wedding Reception Organization (disambiguation) Duty Homeopathy How-to	0.39 0.25 0.23 0.19 0.16 0.13 0.13	54	0.11	54	0.26	60	0.24
09	Tenant's Rights	90	-	-	85	0.29	85	0.29	85	0.29
10	Culture Shocks	45	Culture shock Culture Education Cultural identity Work Autobiography	0.81 0.52 0.44 0.18 0.16 0.13	9	0.00	12	0.00	21	0.05
11	Deep Sea Creatures	16	Marine biology Deep sea creature	0.18 0.17	16	0.15	16	0.14	16	0.11

Continued on Next Page. . .

Table 6 – Continued

ID	Title	R	Wikipedia Concepts		SIM		SIMTR		SIMEXT	
			Title	W	R	MAP	R	MAP	R	MAP
12	Carnivorous Plants	10	Carnivorous plant	0.90	10	1.0	10	0.95	10	1.0
			Carnivore	0.81						
			Plant	0.61						
			Flora	0.45						
			Life	0.16						
13	African Tales	47	Music of Africa	0.70	12	0.01	12	0.01	11	0.01
			Traditional music	0.30						
14	Underground Railways	50	-	-	24	0.07	24	0.07	24	0.07
15	Women in the Middle East	17	Middle East	0.80	9	0.01	9	0.01	13	0.01
			Women's rights	0.18						
			Muslim conquest of Syria	0.18						
16	Rwanda Massacres	21	Rwanda	0.96	21	0.21	21	0.59	21	0.33
			Rwandan Genocide	0.85						
			Genocide	0.76						
			List of events named massacres	0.18						
17	Slavery in Antiquity	18	Slavery in antiquity	0.96	16	0.20	17	0.23	16	0.24
			History of slavery	0.86						
			Slavery	0.85						
			Classical antiquity	0.79						
			Ancient history	0.64						
			History of antisemitism	0.18						
18	Telephony and Telegraphy	49	Invention of the telephone	0.70	46	0.32	46	0.31	46	0.27
			Telegraphy	0.68						
			Telephone	0.59						
			Telephony	0.43						
			Invention	0.14						
19	Healing with Stones	9	The Healing	0.45	3	0.03	3	0.04	3	0.04
20	Digital Photography	53	Digital photography	0.79	52	0.91	52	0.93	52	0.83
			Photography	0.26						
			Digital	0.22						
21	Rock Climbing for Beginners	10	Rock climbing	0.48	9	0.09	9	0.08	9	0.09
22	Irish Saints	39	Saint Patrick	0.88	18	0.05	30	0.15	30	0.13
			Ireland	0.80						
			Saint	0.60						
			Hagiography	0.40						
23	Apes Learning Skills	19	Ape	0.89	10	0.08	15	0.09	15	0.11
			Monkey	0.84						
			Learning	0.12						
24	Albert Einstein	21	Albert einstein	0.95	21	0.55	21	0.46	21	0.54
			Autobiography	0.21						
25	Plant Diseases	212	Cogeneration	0.67	129	0.17	107	0.22	141	0.31
			Plant pathology	0.60						
			Plant	0.60						
			Disease	0.32						
			Treatment	0.22						

Continued on Next Page. . .

Table 6 – Continued

ID	Title	R	Wikipedia Concepts		SIM		SIMTR		SIMEXT	
			Title	W	R	MAP	R	MAP	R	MAP
26	Oil Refining	63	Oil refinery Oil Petroleum industry Geopolitics Refining European Economic Community	0.84 0.80 0.48 0.27 0.25 0.14	34	0.08	40	0.17	43	0.16
27	Female Martyrs	8	Biography	0.14	5	0.05	5	0.07	5	0.06
28	History of the Camera	20	History of the camera Photography Camera Polish language	0.95 0.54 0.33 0.15	19	0.21	19	0.12	19	0.18
29	Garden Shows	44	Garden	0.14	17	0.00	17	0.03	20	0.02
30	Wedding Traditions	90	Wedding Tradition Ceremony	0.54 0.18 0.11	34	0.14	38	0.17	59	0.20
31	Terrace Gardens	83	Windowbox Balcony Imaginary unit	0.36 0.36 0.10	38	0.10	41	0.11	40	0.11
32	Mythology in Contemporary Literature	25	Contemporary literature Mythology Literature	0.46 0.26 0.15	6	0.09	11	0.02	11	0.07
33	Modern Persian Language	61	Persian language Language	0.89 0.45	60	0.12	58	0.48	60	0.41
34	The Normandy Landings	44	Normandy Landings Normandy Invasion of Normandy 1944 Operation Overlord Allied invasion of Sicily Allies of World War II Operation Downfall	0.93 0.89 0.85 0.72 0.67 0.62 0.61 0.21	42	0.65	43	0.65	43	0.63
35	European Educational Systems	235	Education School Student University System	0.50 0.26 0.17 0.12 0.11	60	0.06	70	0.03	81	0.06
36	Urban Parks and Gardens	53	Parks and gardens of Melbourne Urban park Park Garden	0.48 0.47 0.35 0.28	10	0.00	12	0.00	11	0.00
37	Contemporary European Architecture	45	Contemporary architecture Contemporary art Architecture Europe Illustration Photograph Architect	0.52 0.43 0.37 0.33 0.31 0.19 0.14	12	0.00	18	0.05	15	0.01

Continued on Next Page. . .

Table 6 – Continued

ID	Title	R	Wikipedia Concepts		SIM		SIMTR		SIMEXT	
			Title	W	R	MAP	R	MAP	R	MAP
38	Natural History Museums	106	Museum Natural History List of natural history museums History	0.97 0.83 0.50 0.12	26	0.07	82	0.26	71	0.30
39	Ozone Depletion	44	Ozone depletion Ozone Stratosphere Polar region Earth Chemical polarity Depletion	0.96 0.85 0.76 0.54 0.37 0.16 0.15	43	0.51	43	0.62	43	0.54
40	European Union Labour Laws	23	European Union Labour law Regulation Employment Occupational safety and health Trade union Labour Party (UK) Government	0.97 0.51 0.46 0.36 0.33 0.29 0.22 0.12	10	0.01	14	0.01	15	0.05
41	Sailing for Beginners	12	Boating Sailing	0.48 0.37	11	0.10	11	0.12	11	0.08
42	Decorating Children's Rooms	17	How-to Decoration Child	0.32 0.17 0.16	11	0.16	12	0.17	12	0.14
43	Women Spies	8	-	-	4	0.03	4	0.03	4	0.03
44	Aztec Religions and Myths.	22	Aztec Aztec mythology Mythology Religion Major religious groups	0.89 0.84 0.62 0.58 0.21	12	0.13	14	0.37	16	0.39
45	Traditional Costumes	65	National costume Tradition Dress	0.47 0.29 0.12	29	0.07	31	0.03	38	0.09
46	European Regional Development	35	European Union Europe Regional development Region	0.74 0.59 0.38 0.11	5	0.00	16	0.01	20	0.03
47	Maths for Children	185	-	-	105	0.28	105	0.28	105	0.28
48	Knitting for Children	13	Knitting How-to	0.59 0.12	13	0.07	13	0.09	13	0.10

Continued on Next Page. . .

Table 6 – Continued

ID	Title	R	Wikipedia Concepts		SIM		SIMTR		SIMEXT	
			Title	W	R	MAP	R	MAP	R	MAP
49	Human Gene Manipulation	43	Gene	0.82	41	0.34	41	0.27	41	0.37
			Genetic engineering	0.49						
			Morality	0.44						
			Human anatomy	0.32						
			Human body	0.31						
			Genetics	0.25						
			Human	0.23						
			Research	0.23						
			Manipulation	0.23						
			Ethics	0.20						
			Body	0.14						
50	Contemporary French Philosophers	30	Philosophy	0.48	9	0.00	22	0.08	23	0.12
			French philosophy	0.11						

References

[1] Eneko Agirre, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. Clef 2008: Ad hoc track overview. In *Proceedings of the CLEF 2008: Workshop on Cross-Language Information Retrieval and Evaluation*, Aarhus, Denmark, 2008. 4, 8

[2] A. AleAhmad, E. Kamalloo, A. Zareh, M. Rahgozar, and F. Oroumchian. Cross language experiments at persian@clef 2008. In *Proceedings of the CLEF 2008: Workshop on Cross-Language Information Retrieval and Evaluation*, Aarhus, Denmark, 2008. CLEF 2008 Organizing Committee. 9

[3] James P. Callan, W. Bruce Croft, and John Broglio. Trec and tipster experiments with inquery. *Inf. Process. Manage.*, 31(3):327–343, 1995. 2

[4] Jon Dehdari and Deryle Lonsdale. A link grammar parser for Persian. In Simin Karimi, Vida Samiian, and Don Stilo, editors, *Aspects of Iranian Linguistics*, volume 1. Cambridge Scholars Press, 2008. 9

[5] Amir Hossein Jadidinejad and Fariborz Mahmoudi. Qiau at clef2009: Persian track. In *Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation*, Corfu, Greece, September 2009. CLEF 2009 Organizing Committee.

[6] Simin Karimi. Persian or farsi? `http://www.u.arizona.edu/~karimi/Persian%20or%20Farsi.pdf`. 8

[7] R. Karimpour, A. Ghorbani, A. Pishdad, M. Mohtarami, A. AleAhmad, and Amiri A. Using part of speech tagging in persian information retrieval. In *Proceedings of the CLEF 2008: Workshop on Cross-Language Information Retrieval and Evaluation*, Aarhus, Denmark, 2008. CLEF 2008 Organizing Committee. 9

[8] Donald Metzler and W. Bruce Croft. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40(5):735–750, 2004. 2, 3, 4, 6

[9] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM. 2, 3

[10] David Milne and Ian H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM. 2, 3

[11] Paul Ogilvie and Jamie Callan. Experiments using the lemur toolkit. In *In Proceedings of the Tenth Text Retrieval Conference (TREC-10*, pages 103–108, 2002. 6

[12] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. Indri: A language-model based search engine for complex queries (extended version). IR 407, University of Massachusetts, 2005. 2, 3, 4, 6

[13] Masoud Tashakori, Mohammad Reza Meybodi, and Farhad Oroumchian. Bon: The persian stemmer. In *EurAsia-ICT*, pages 487–494, 2002. 9

[14] Ian H. Witten and David Milne. An open-source toolkit for mining Wikipedia. In *(to appear)*, 2009. 3