World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:1, 2008

Language and Retrieval Accuracy

Ahmed Abdelali, Jim Cowie, and Hamdy S. Soliman

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

***Abstract*—**One of the major challenges in the Information Retrieval field is handling the massive amount of information available to Internet users. Existing ranking techniques and strategies that govern the retrieval process fall short of expected accuracy. Often relevant documents are buried deep in the list of documents returned by the search engine. In order to improve retrieval accuracy we examine the issue of language effect on the retrieval process. Then, we propose a solution for a more biased, user-centric relevance for retrieved data. The results demonstrate that using indices based on variations of the same language enhances the accuracy of search engines for individual users.

***Keywords*—**Information Search and Retrieval, Language Variants, Search Engine, Retrieval Accuracy.

I. Introduction

THE Internet continues to grow dramatically and more and more relevant information should be available on any topic. However, the amount of information available to Internet users may often block them finding the data they need in a timely fashion. The ability to retrieve appropriate data seems to be driven by other factors than relevance. Paid advertisements and ranking methodologies are influencing, if not governing, the retrieval process. It is difficult to find a relevant document because often it is way down in the list of retrieved documents. Surprisingly, an Internet user might see documents on the top of the list from Australia or New Zealand while he or she was looking for a business a few miles away from his or her door in Los Angeles. The luxury of having unlimited access to information which is not obstructed by geographical borders now is the origin of another issue – "How can a user make a search which is focused on their needs and knowledge".

To locate appropriate information Internet users intuitively use their knowledge and background to craft their queries [1]. An English speaker from the UK may say "tube light" while a US speaker will refer to the same item as a "fluorescent lamp." One user knows the term "stroller", the other "baby buggy." Typically, Internet users would prefer to be given all relevant documents ranked in accordance with their own perspective and their own biases [30,12,14,31]. Increasingly, it is becoming necessary to provide users with new solutions to make their searches more efficient, convenient, and personal, which will allow exploring the internet efficiently [22]. Users can manipulate current systems by biasing their query by adding terms with the intention of indicating the specific dialect of language they are using. Such additions may not have the intended effect, since it could be related to an advertisement or match up with a word from the metakeys in the target website. For example, we added the word "USA" to the query "cooker" in a query to the Google search engine to indicate the "American" English meaning. The results obtained still contain sites from UK – ranked in the fourth position of the top ten retrieved documents. Other dialect specific words could be used as well, like "color" instead of "colour".

Advances in language identification [7,13,15,21] make it possible to accurately identify one language (or even dialect) from among others. By training a different language model for each language, the conditional probability for a new sentence can be computed for each language model. These probabilities for each language can be compared to make a prediction. Cavnar and Trenkle [7] tested a language classification system that achieved 99.8% correct classification. McNamee [25] described a system to identify language using data obtained from the World Wide Web, which achieved an accuracy approaching 100% on a test comprised of ten European languages. Using the same approach, variations in one language could be highlighted, especially if there are enough lexical or syntactical differences. Such features help identify the user's language, which improves our notions about the answers that might be expected.

To explore the extent of this problem, we carried out experiments on a number of the widely used search engines to demonstrate the effect of the users' dialect on the retrieval process. For the same purpose we also experimented with different English corpora of the same language from different regions to assess the behavior of queries and whether or not they would retrieve relevant documents from across regions. We demonstrated that queries mostly retrieve local documents from the same language class, which implies that for better retrieval quality search engines need to consider a query's origin. Such considerations will help to narrow down the search and provide more accurate answers. Our approach is to pre-classify the websites to be indexed into classes based on language variation. An incoming query will trigger a search in its appropriate language class—instead of the entire database—speeding up obtaining more accurate results. This is confirmed by our experimental results. We use English as

Manuscript received May 27, 2008.

Ahmed Abdelali is with the Physical Science Laboratory, MSC 3PSL, New Mexico State University, Las Cruces NM 88003 USA (phone: 575-646-5711; fax: 575-646-6218; e-mail: aabdelali@psl.nmsu.edu).

Jim Cowie is with the Psychology Department, MSC 3452, New Mexico State University PO Box 30001Las Cruces, NM 88003-3452 USA (e-mail: jcowie@nmsu.edu).

Hamdy S. Soliman is with the Computer Science Department, New Mexico Institute of Mining and Technology. Socorro, NM 87801-0389, USA (e-mail: hss@nmt.edu).

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

our case study because of its wide dominance in the World Wide Web and the availability of test data.

In the section two (below) we explore the problem by using sample queries on the most popular search engines. In section three we summarize technologies used for language identification. In section four we present research on English corpora and efforts related to the research area. Section five presents our research hypothesis. In section six we present our experiments and discuss our results. Finally, in section seven we conclude with recommendations to improve current search engines.

II. Exploring the Problem

To gain a better understanding of the problem, we need to study the performance of different 'languages' queries on search using existing search engines. We used three users based in the USA, Ireland (IR), and France (FR). They used the same queries on four popular search engines—Alltheweb, Google, MSN, and Yahoo. The queries were constructed using words that are known to carry more than one meaning in dialects of English (British English, American English, and Indian English). The results were that each user received almost identical answers. This is shown in Tables V and VI. These results are puzzling and hence not satisfactory to these three different users. A user would prefer to retrieve a list of documents ordered based on his/her needs (recognizing clearly his/her language biases). The small differences between the results shown in Tables IV and V could be related simply to the fact that the users submitted their queries at different times, and that not all the queries were tested on the same day.

Knowing that "cooker" in the UK is used to mean what is referred to in the US as a "stove," means it is unnecessary and misleading to show the US user a list of stores in the UK that supply stoves. Our other query exhibited a similar phenomenon the case of "pavement" in UK parlance being equivalent to "sidewalk" in USA usage. (See Tables V-VI in Appendix A). The results returned by the search engines contains contradicting answers and are mostly inconclusive for the fact that the results were from websites of different regions and they provide different meanings and concepts. In Tables I and II, we tried to use standard measurement (recall-precision metric) to present the issue in a standard way to gain more understanding of the problem. For the case of US and IR user, the assumption is that any website not from their area will not be relevant, but the issue gets more complicated when considering the FR user. Ideally the retrieval system will favor local results if none then the results from the closer locality will be more relevant. Considering such analogy for an FR user, results from UK would be more relevant than results from US or Canada.

TABLE I
Recall Results for the Queries "Pavement Width"

Recall at	US	IR	FR
5	1.00	0.00	0.00
10	0.80	0.00	0.00
15	0.80	0.00	0.00
20	0.85	0.00	0.00

TABLE II
Recall Results for the Queries "Cooker Price"

Recall at	US	IR	FR
5	0.20	0.60	0.60
10	0.20	0.70	0.50
15	0.33	0.67	0.60
20	0.40	0.60	0.55

The problem arises when a word in a query carries different meanings in geographical areas with different variations of the same language. The ultimate solution would be to disambiguate any user query. There is, however, no definite answer for this problem; it is more practical to provide a solution through the exclusion of answers that do not fit a user's needs and preferences. The phenomenon is not unique to English, but the same issue was observed in other languages as well [2].

Appendix A provides more details about the queries, the settings used in the experiments, excerpts from the documents retrieved, regions, and URLs.

Word-sense disambiguation (WSD) attempted to address the same issue by attempting to assign one of several possible senses to a particular occurrence of a word in text [4], the proposed approach could be an initial step of dividing language into sub-classes that share a set of meaning. Later WSD would identify each of the possible senses benefiting from the narrowed possibilities for the meanings. Other approaches that rely on Click-Trough logs are among the newest attempts to identify user behavior and preferences and hence provide more satisfactory results [3,16].

III. Identification of Language Differences

Intuitively, users of a search engine use their native language to formulate queries, and they expect the retrieved documents to be in the same language. However, for the most commonly used languages there are large divergences based on hosting countries and regions, despite the existence (in some cases) of standardization bodies. Identifying these differences lets us potentially improve searching and also allows us to avoid confusion in the interpretation of documents.

Researchers have been interested in the problem of language identification [7,13,15,21] and have explored the issue from different perspectives. One of the major approaches is based on statistical language modeling.

Statistical models are usually built by collecting statistics from a large set of data. For a textual data, this can be done on the word or character level. While Grefenstette [15] compared

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

common words and common trigrams using a zeroth-order Markov model based on words and trigrams. Dunning [13] made a more exhaustive comparison using models of order zero-through six on characters (i.e., from single letters to sequences of seven letters); he also found that trigrams work well. Cavnar and Trenkle [7] tested an n-gram text categorization system on a collection of Usenet newsgroup articles written in different languages and a set of articles from different computer oriented newsgroups. The language classification system achieved 99.8% correct classification. McNamee [25] described a system to identify language using data obtained from the World Wide Web that achieved an accuracy approaching 100% on a test comprising ten European languages. Cowie, Yevgeny and Zacharski [10] described a language recognition algorithm for multilingual documents based on mixed order n-grams, Markov chains, and maximum likelihood. In addition to language recognition, their proposed algorithm has a separate verification step that assures, with a controllable degree of certainty, that the text to be classified is actually in the closest language. Their algorithm was tested on 34 languages.

Studying variants within a language using text was not widely addressed if not at all. On the other hand attempts to study dialects using speech and acoustic features were made since few decades [17,28,8,32].

IV. Corpora and Corpus-based Analysis: The Case of English

Using text collections (corpora) in language study is not a new idea. Since early times, collecting word lists and identifying context in particular texts has been a legitimate linguistic enterprise. Other productions of corpus analysis are lists of most frequent words from single texts or from collections of texts, statistics on co-ocurrence and a wide variety of analyses on syntax and semantics. Most dictionary work in English is now grounded in large, evolving corpora. Areas of studies where corpora play an important role include language acquisition, syntax, semantics, and comparative linguistics. “Even if the term 'corpus linguistics' was not used, much of the work was similar to the kind of corpus based research we do today with one great exception—they did not use computers.” (W3 Corpus Tutorial – University of Essex) [33].

Modern corpora combine text sampled from different areas and genres. The first were the Brown Corpus—the first modern, electronically readable corpus of Standard American English and the LOB (Lancaster Oslo Bergen) Corpus of British English. The former corpus consists of one million words of American English texts printed in 1961. The texts were sampled in different proportions from 15 different text categories: Press (reportage, editorial, reviews), Skills and Hobbies, Religion, Learned/Scientific, Fiction (various subcategories), etc. Corpora available nowadays include:

- The LOB, Lancaster-Oslo-Bergen, corpus (British English)
- The Kolhapur Corpus (Indian English)
- The London-Lund Corpus of Spoken British English (LLC)
- The British National Corpus (BNC)
- The American National Corpus (ANC)

The American National Corpus (ANC) initially proposed at the first LREC in 1998 [18], is now well on its way to realization. The first data in its base level representation was scheduled to be available to the NLP community and consortium members by 2004.

The First Release of ANC data is now available from the Linguistic Data Consortium (LDC). The First Release consists of over 10 million words marked for part of speech and lemma, in both a "stand-off" and merged format. The Second Release of ANC data released by LDC on Dec. 2005 contains over 20 million words; an additional 10 million words added on top of the First Release. The full corpus consists of (at least) 100 million words annotated for part of speech, together with search and retrieval software.

Recently smaller corpora of English have been released by The International Corpus of English (ICE), an institution with the primary goal of collecting material for comparative studies of English worldwide. The corpora were compiled from different English speaking areas among which are India, the Philippines, Singapore – Fig. 1, and East Africa.

And when collecting the papers and this is something that I	**raised**	
Now the points have been	**raised**	by my learned friend was that
they have	**raised**	triable issues for it to be tried in the full court
invalid or in any event the Defendants have	**raised**	an arguable case for
It added that he	**raised**	the idea in a letter to leaders of the
war was not yet over had	**raised**	concern about the future of the UN peace
The top rate on personal income too has been	**raised**	by five per cent
In Vietnam, economic reforms have	**raised**	expectations that probably
to inform you that it has	**raised**	sufficient funds to cover a major part of
With his mother at work, the children were left to be	**raised**	mostly by
Sui	**raised**	her eyes from the ground without any idea
He could even hear the	**raised**	voice of a mother furiously and nasally

Fig. 1 Excerpts from ICE-Singapore

Although Kennedy [19] reported that "non-trivial syntactic differences between British and American English have been notably harder to find in corpus-based studies," Wulff et al. [34] concluded that over the last two decades, the growing availability of variety-specific corpora has enormously enhanced the study of regional varieties of English, triggering studies that cover a broad range of aspects such as vocabulary use, conversational style, or the use of modal verbs. Fig. 2 and 3 show examples of vocabulary use for word "buggy" between American and British English. The examples have been collected from Collins WordbanksOnline English corpus of modern written and spoken English on which many Collins English dictionaries are based.

is by electric cart or by horse and **buggy**. Trips can be arranged from Shelter Cove
bowl, fish, play tennis, or take a horse and **buggy** ride. Children can participate in the Youth
made, wif" yuh own money an' yuh own li'l love **buggy**?" He pulled on Miguel's coat sleeve. `You don'
Sunday afternoons driving in the yellow-wheeled **buggy** and the matched team of bays from the livery
passed on Sunday afternoon in the glittering **buggy**, Miss Emily with her head high and Homer
refusing to yield its old-time horse-and-**buggy** splendor to the age of the automobile, the
left it on the front porch and returned in his **buggy** to Jasper, but she never got the watermelon,
a strong streak of old-fashioned horse-and-**buggy** agrarianism. Farming, from this perspective,
roar of the automobile disturbed the horse-and-**buggy** atmosphere. So Ford had the state highway
phone-company switching systems are horse-and-**buggy** operations

Fig. 1 "Buggy" from American books, ephemera and radio, from the Collins Wordbanks*Online* English corpus

designed all-terrain vehicle, called a Tundra **Buggy**. Unique to this area, the Caribou is fully
and arctic foxes are frequently seen on the **Buggy** tours and nearly 150 different birds
Being carted around in my pushchair, or Baby **Buggy** as the sticker proudly boasts, makes me an
fro your baby when he's out and about in **buggy** or car seat or in his highchair. [p] [h]
she'll no longer fit into a standard care seat, **buggy** or highchair. Here are some solutions parents
models with padded side bars. [p] Use a lieback **buggy**, supporting your baby with pillows or
other first-time mums too. [p] 1 Choosing a **buggy** can be a nightmare as there are so many
to consider before you buy are whether the **buggy** is safe, solid and has an easy folding
mums in our group discovered that her baby's **buggy** wouldn't fit in her small hatchback unless
down. You should also check if you can fold the **buggy** without having to remove the raincover. This
to thread their way past a gypsies' horse and **buggy** race on the dual carriageway course. [p]
of corridors, but you may be able to take your **buggy** up to the boarding gate before it's stowed
bootees and socks to sling over the side of the **buggy**. A universal movement-freedom for feet! Haven'
he wakes up at night so he pisses in the baby **buggy**. Be glad that it hasn't happened to you-IM
Sputnik, Transvision Vamp and Fabulous, here's **BUGGY** HAIR on come HUGGY BEAR) [p]
s steamer. [p] So she spins away in her wheeled **buggy**, the ricksha-men's dark legs churning up the
teeth and poor skin wheeling two children in a **buggy** and puffing wearily on a cigarette an
last visits, Mrs Merton, seated in her electric **buggy**, led me to a doorway saying: `Come and see my
of a flamingo. Oronte becomes Covington (Niall **Buggy**), an ingratiating and stage-struck drama
difference. [p] [h] Tory MP trapped child in **buggy**;Sir Nicholas Scott banned for year after [/h]
in his pushchair. The boy was caught in the **buggy** between the Volvo estate and a Jaguar. [p]
is out - neighbours have spotted the baby **buggy** outside the luxury Queen Anne mansion on the
The attacker hauled the three-year-old out of a **buggy** by her hair then pushed hysterical Lynda Wall
and her son, aged 11 months, was also in the **buggy**. [p] She was treated for bruising after the
[p] After 10 minutes in hiding an airline **buggy** was summoned and Naomi hopped on board for
be a particular nightmare with a toddler in a **buggy** [p] Large warehouse-style stores on the

Fig. 2 "Buggy" from British books, ephemera, radio, newspapers, and magazines, from the Collins Wordbanks*Online* English corpus

V. Language Differences and their Impact on the Retrieval Process

A. Language Modeling and the Retrieval Process

Statistical Language Modeling (SLM) extends Shannon's noisy channel theory by introducing probability theory to address the question of the rate of information that can be transferred over a noisy channel. Shannon's noisy channel theory states "*a communication channel is a system in which the output depends statistically on the input. It is characterized by a conditional distribution* ***p(y|x)*** *that* ***y*** *emerges from the channel given that* ***x*** *was input*." [21]. In this setting, language modeling has been seen as equivalent to the theory that a word sequence W in text is generated by some source with probability $P(W)$ and transmitted through a noisy channel that transforms the intended W to the observation A with probability $P(A|W)$. This formalism gives us the ability to compare language models. Using the quantity entropy, one can measure and estimate how good a model might be. Ideally, the best fit model should lead to the lowest recognition error rate.

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

To do this we need to compute the average log probability on a word basis for a piece of new text not used in building the model. Let us denote by ***p*** the probability distribution of a segment of text **x** of **k** words. The entropy is defined as:

$$H = \lim_{n\to\infty} -\frac{1}{k}\sum_{k} p(x)\log_2 p(x)$$

or simply

$$H \approx -\frac{1}{n}\log_2 p(x)$$

The idea of entropy is a measure of our uncertainty; the more we know about something the lower the entropy will be because we are less surprised by the outcome of our trial. In the speech recognition community, people tend to refer to perplexity rather than entropy [24]. The relation between the perplexity and entropy is:

$$PP = 2^H$$

Alternatively, using the SLM approach, Ponte and Croft [27] proposed LM for IR. LM can estimate the relevance of a document *d* with respect to a query *q*, by computing the likelihood of generating *q* from *d*, i.e. *p(q|d)*. Cronen-Townsend et al. [11] showed that among the problems that contribute to the failure of retrieval systems is one that arises when the language models for the document and the query are different. They introduced the "clarity score," which is simply the relative entropy between the query and collection language models. The score can estimate the divergence between the query and the documents.

$$\text{clarity score} = \sum_{w\in V} P(w\,|\,Q)\log_2 \frac{P(w\,|\,Q)}{P_{coll}(w)}$$

This score will help to predict the performance of the query without relevance information. Cronen-Townsend et al. [11] concluded that there is a strong correlation between the clarity score of a test query with respect to the appropriate test collection and the performance of the retrieval system. Hence, our approach aims to obtain a high clarity score by clustering all the indexed documents into classes, based on language variation, facilitating a direct query access to a more precise class of documents. If the language model is successful it will assign a high probability to this test text, with the result that the language model will have a low perplexity [9].

On the same theme, i.e minimizing the perplexity, Sethy, Georgiou, and Narayanan [29] used a set of seed documents to generate queries that are able to retrieve and collect from the web documents relevant to a specific "topic." The seed documents ensure the consistency and relevancy of the collected documents. In general, a query from a different domain will not perform as well as a query from the same domain or topic. Azzopardi et. al. [5] explored the relationship between the language model perplexity and IR precision-recall. Azzopardi et. al. [6] concluded that "*if we consider that the topic based approach is in fact a novel implementation of the Cluster Hypothesis within the Language Modeling framework, then we have provided empirical evidence that shows using the inherent topical structure can achieve improved IR performance*."

If we extended the notion of "topic" to be a "geographical region," a native user in a region formulates a query that retrieves more relevant documents than a non-native user. Therefore, if relevant documents exist, there is more potential that the queries crafted by a native user will retrieve them.

B. Suggested Approach

The revised scheme of our novel IR system will be described at the following three levels:

1) Classes Construction Level:
   This step prepares the corpora that will be used for identification and classification. After comparing the perplexities (entropies) of different corpora, the corpora will be clustered, for optimal future document classification. Such a task can be carried out using any clustering algorithm. In our case, we used the k-means algorithm [23].
   Other clustering algorithms could be also used in this task, such as Kohonen Feature Map [20] and PCA [35].
   Each of the resulting clusters will define a new corpus. All the corpora in one cluster will be merged to build one LM. The new models will serve as the "language identifier" (see Figs. 4 and 5).
2) Indexing Level:
   Once the classes are determined, for every spidered document, the perplexity of the document and the LMs will be computed. The document will be assigned to the cluster (class) with the minimum perplexity. In order to ensure coverage and representativeness, the LM can be updated incrementally by including recent documents that belong to the same cluster and dumping out older documents, in the above process (see Fig. 4).
3) Querying Level:
   The classes' LMs provide the likelihood of a query "q" being generated from a document "d" from a certain class. Therefore, the "language identifier", in Fig. 4, will find the index of the closest cluster (of corpora) to "q". All documents relevant to the input query will be retrieved from the cluster with the obtained index, i.e., the selected language variant.
   In an interactive environment, e.g. the Internet, a user could elect to explore wider ranges of alternatives. For such purpose, a user would provide information which can be combined in the index search, user IP address (Internet Protocol address is a unique global addressing amongst computers) which can validate the outcome of the model. The user can choose to override such a classification and query all the indexes, in the same fashion as is done by current search engines.
   In our model the process of document classification will be done off-line (before querying the system) while

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

populating the search database (see Fig. 5).

To evaluate the above mechanism, we used the ICE corpora from East Africa, India, Singapore, and the Philippines to build language models and test our approach. One of our research hypothesis is that given a document from one region, the document perplexity with the region's LM will be the smallest compared to other regions' LMs.

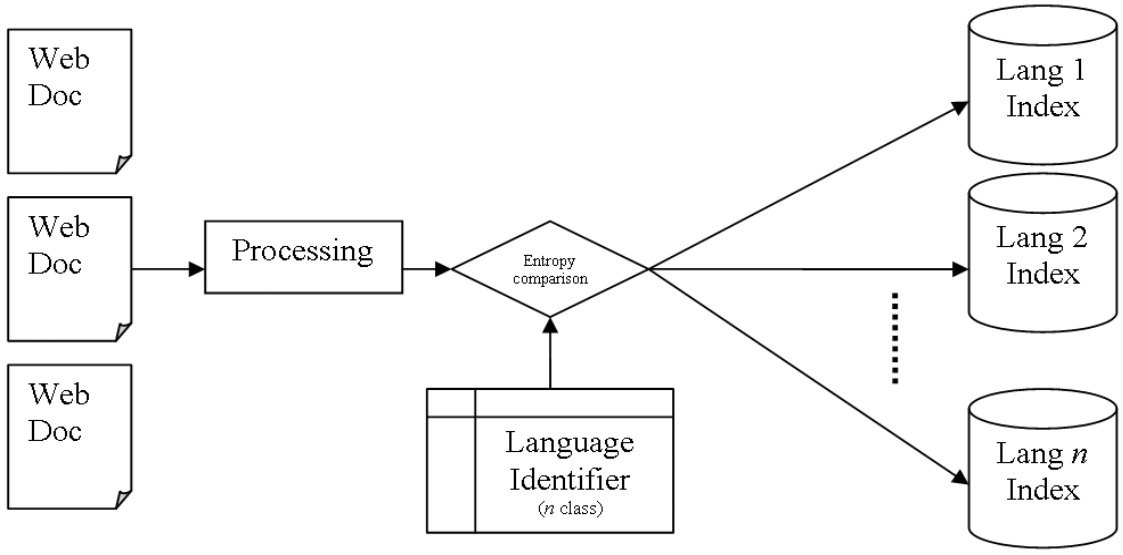

Fig. 3 Process of classification for crawled internet pages

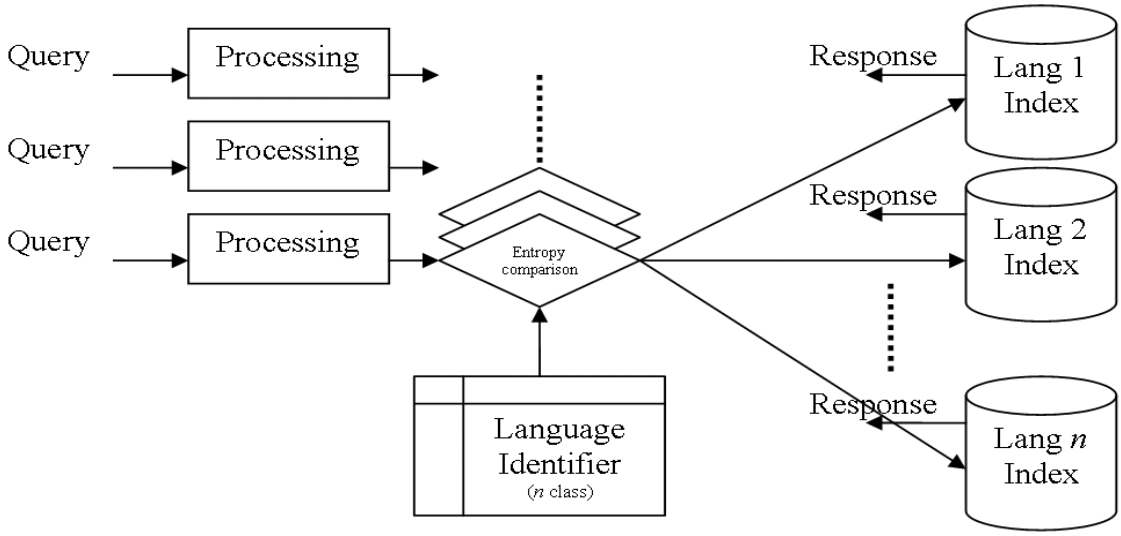

Fig. 4 Process of searching on a multi-variant indices engine

VI. Evaluation

We used four published corpora of English from the International Corpus of English (ICE), an institution whose primary goal is collecting material for comparative studies of English worldwide. The collections were selected from countries where English is the official language beside other native languages. Table III provides information about the collections.

To check the validity of our hypothesis, we randomly divided each corpus into ten parts. We used one-tenth for building a test set and the remaining nine-tenths to contribute to the formation of the final model language classes. Then, we computed the entropies (perplexity could be used instead) for each incoming document from the test set with respect to the language model built from the remaining 9/10 of each of the four corpora. The lower the value of the entropy indicates that the test text is more similar to the language model text. A validity check was run to make sure that the minimal entropy was obtained from the corpus where the test document had been selected initially. We repeated the above process while varying the choice of the one-tenth from each corpus to cover

all the possible combinations of the ten divisions of each corpus resulting in ten different runs, as shown in Table IV. The results show the average entropies for the test samples for each model. For all ten experiments the model identified correctly 88.75% of the documents in the test sets, which demonstrates its ability to discriminate documents from different regions. Cases where the models failed to identify the language correctly are indicated in bold face fonts.

The results of Table IV validate the hypothesis, which implies that relevant documents for a query will be within the same set of language models—classes—that share lower entropy. This could explain the failure in Tables V and VI where the UK users did not retrieve relevant results based on their biases. The lack of a classified search based on language variations in existing search engines could be the cause.

TABLE III
ENGLISH CORPORA DETAILS

Corpus	Source	N. of Docs	Size	N. Of Words	Unique Words
ICE-EA	East Africa	500	7816K	1359849	35930
ICE-Ind	India	500	6176k	1105959	39156
ICE-Phi	Philippines	478	6192k	1123396	37700
ICE-Sin	Singapore	500	6096k	1099433	33357

TABLE IV
ENTROPY ASSESSMENT OF DIFFERENT CORPORA

Entropy	RUN0	RUN1	RUN2	RUN3	RUN4	RUN5	RUN6	RUN7	RUN8	RUN9
ICE-EA Language Model										
ICE-EA	10.586	11.258	10.814	10.76	11.34	11.346	11.732	11.262	11.59	11.22182
ICE-Ind	10.669	**10.587**	11.09122	11.13429	**11.21837**	11.51286	**11.68306**	12.26816	11.99122	11.76414
ICE-Phi	10.869	**10.994**	11.076	11.3816	11.3728	11.7716	**11.6534**	12.252	11.9352	11.7832
ICE-Sin	10.790	**10.721**	11.0754	11.2728	11.0014	11.4168	**11.526**	12.0162	11.692	11.6168
ICE-Ind Language Model										
ICE-EA	10.902	11.592	11.048	**10.664**	11.726	11.726	12.118	**11.816**	**11.954**	**11.64727**
ICE-Ind	10.240	10.294	10.96082	10.81673	11.01204	11.61449	11.16408	12.09245	11.98796	11.75103
ICE-Phi	10.954	11.092	11.0936	11.3778	11.4508	12.1138	11.8392	12.496	12.2688	12.0534
ICE-Sin	10.563	10.647	11.1396	11.1644	11.0798	11.768	11.6892	12.302	**11.9064**	11.8444
ICE-Phi Language Model										
ICE-EA	10.904	11.668	11.092	**10.62**	11.77	11.864	12.238	**11.916**	12.016	**11.67727**
ICE-Ind	10.508	10.543	11.23469	11.24245	11.34388	11.91449	11.90204	12.58102	12.26878	12.03466
ICE-Phi	9.810	9.894	10.4942	10.7788	10.9044	11.7356	11.4862	12.3332	11.873	11.6794
ICE-Sin	10.512	10.576	11.1002	11.2154	11.1502	11.7562	11.7896	12.3992	11.9816	11.9204
ICE-Sin Language Model										
ICE-EA	10.856	11.578	11.002	**10.55**	11.7	11.776	12.208	**11.892**	11.95	**11.61455**
ICE-Ind	10.436	10.464	11.14041	11.13265	11.27122	11.92551	11.88184	12.47592	12.19918	11.85759
ICE-Phi	10.796	10.923	11.062	11.3506	11.3986	12.1358	11.7726	12.541	12.256	11.9464
ICE-Sin	10.081	10.156	10.8274	10.966	10.8064	11.5918	11.5886	12.2874	11.7604	11.659

VII. CONCLUSIONS AND FUTURE WORK

In this paper we showed that current search engines that ignore query language source will suffer from a lack of accuracy in the retrieved documents. Users might get irrelevant retrieved documents according to their biases. Hence we introduced a more accurate model for document retrieval that divides the search data into classes of related language variation. Experimental results showed significant precision for locating the correct document which promises a significant enhancement in the accuracy of retrieved documents.

The availability of more corpora from other languages will advance the understanding of the above problem and the development of appropriate solutions. One way to achieve such goals is that instead of statically obtaining a pre-prepared class, we dynamically manufacture classes, as we collect documents from different sites, via the application of the ART neural model.

The release of corpora similar to the ANC, in kind or size, will allow the quantification of language differences and their impacts. For future work, we intend to investigate the language variations in user submitted queries, via query logs.

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

APPENDIX

TABLE V
QUERY "PAVEMENT WIDTH" USING "GOOGLE" SEARCH ENGINE

Rank	US	Ireland	France
1.	www.metrocouncil.org	www.metrocouncil.org	www.pacode.com
2.	www.pacode.com	www.pacode.com	www.pacode.com
3.	www.pacode.com	www.pacode.com	www.dot.wisconsin.gov
4.	www.dot.wisconsin.gov	www.dot.wisconsin.gov	www.metrocouncil.org
5.	www.cityofdover.com	www.dot.state.mn.us	www.cityofdover.com
6.	www.toronto.ca	www.cityofdover.com	www.dot.state.mn.us
7.	nemo.uconn.edu	www.toronto.ca	www.toronto.ca
8.	www.northlibertyiowa.org	nemo.uconn.edu	nemo.uconn.edu
9.	www.mincad.com.au	www.mincad.com.au	www.northlibertyiowa.org
10.	www.fhwa.dot.gov	www.northlibertyiowa.org	www.naperville.il.us
11.	www.naperville.il.us	www.fhwa.dot.gov	www.mincad.com.au
12.	www.afcee.brooks.af.mil	www.naperville.il.us	www.kytc.state.ky.us
13.	www.toronto.ca	www.usace.armu.mil	www.fhwa.dot.gov
14.	ppc.uiowa.edu	www.afcee.brooks.af.mil	ppc.uiowa.edu
15.	www.co.gwinnett.ga.us	www.toronto.ca	www.dot.state.mn.us
16.	www.usace.army.mil	www.cctrail.org	www.dpac.tas.wa.us
17.	www.cctrail.org	www.lakecitygov.com	www.ci.seattle.wa.us
18.	www.plannersweb.com	www.ci.seattle.wa.us	www.city.vancouver.bc.ca
19.	www.tfhrc.gov	www.tfhrc.gov	www.afcee.brooks.af.mil
20.	www.co.honolulu.hi.us	www.ata.com	www.co.honolulu.hi.us

TABLE VI
QUERY "COOKER PRICE" USING "MSN" SEARCH ENGINE

Rank	US	Ireland	France
1.	www.texasirons.com	www.texasirons.com	www.aga-cooker-style.co.uk
2.	www.asiachi.com	www.asiachi.com	www.aga-cooker-style.co.uk
3.	www.chefsresource.com	www.chefsresource.com	www.pigroast.com
4.	virat.8m.com	virat.8m.com	www.pressure-cookers.info
5.	www.pressure-cookers.info	www.pressure-cookers.info	www.pressure-cookers.info
6.	www.homeandgifts.schwans.com	www.homeandgifts.schwans.com	www.amazon.com
7.	www.homeandgifts.schwans.com	www.homeandgifts.schwans.com	virat.8m.com
8.	www.kck.com	www.kck.com	shop.allrecipes.com
9.	www.popularelect.com	www.popularelect.com	www.hearthware.com
10.	www.asseenontv.com	www.asseenontv.com	www.kck.com
11.	www.2sale.us	www.2sale.us	www.asiachi.com
12.	www.rompbklyn.com	www.rompbklyn.com	www.cdbaby.com
13.	www.companyscoming.com	www.companyscoming.com	www.ethicalcookshop.co.uk
14.	www.cdbaby.com	www.cdbaby.com	www.fagorpressurecookers.com
15.	shop.allrecipes.com	shop.allrecipes.com	www.asseenontv.com
16.	www.barbecue-store.com	www.barbecue-store.com	www.pigroast.com
17.	www.grilllovers.com	www.grilllovers.com	www.pressurecooker.com.au
18.	missvickie.com	missvickie.com	www.pressurecooker.com.au
19.			www.amazon.co.uk
20.			cooker-hoods.pagedeals.com

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

TABLES VII, VIII
ORIGINS OF DOCUMENT RETRIEVED FOR QUERY "PAVEMENT WIDTH" AND "COOKER PRICE"

Query: pavement width			
Rank	US User	IR User	FR User
1.	USA	USA	USA
2.	USA	USA	USA
3.	USA	USA	USA
4.	USA	USA	USA
5.	USA	USA	USA
6.	Canada	USA	USA
7.	USA	Canada	Canada
8.	USA	USA	USA
9.	Australia	Australia	USA
10.	USA	USA	USA
11.	USA	USA	Australia
12.	USA	USA	USA
13.	Canada	USA	USA
14.	USA	USA	USA
15.	USA	Canada	USA
16.	USA	USA	USA
17.	USA	USA	USA
18.	USA	USA	Canada
19.	USA	USA	USA
20.	USA	USA	USA

Query: cooker price			
Rank	US User	IR User	FR User
1.	UK	UK	UK
2.	UK	UK	UK
3.	USA	USA	USA
4.	UK	USA	USA
5.	UK	UK	UK
6.	USA	UK	UK
7.	UK	UK	USA
8.	UK	USA	UK
9.	UK	UK	USA
10.	UK	UK	USA
11.	USA	USA	UK
12.	UK	USA	UK
13.	USA	UK	UK
14.	USA	UK	USA
15.	UK	UK	UK
16.	USA	UK	UK
17.	USA	USA	USA
18.	USA	USA	USA
19.	UK	USA	UK
20.	UK	UK	USA

TABLE IX
EXCERPT FROM DOCUMENTS RETRIEVED BY THE QUERY "PAVEMENT WIDTH"

Rank	Website	Origin	Comment
1.	www.metrocouncil.org	USA	
2.	www.pacode.com	USA	
3.	www.pacode.com	USA	
4.	www.dot.wisconsin.gov	USA	
5.	www.cityofdover.com	USA	
6.	www.toronto.ca	Canada	**Pavement width** required to provide minimum 4.0 m wide clearance
7.	nemo.uconn.edu	USA	
8.	www.northlibertyiowa.org	USA	Local and industrial streets will have a minimum **pavement width** of 29 feet.
9.	www.mincad.com.au	USA	
10.	www.fhwa.dot.gov	USA	
11.	www.naperville.il.us	USA	The minimum **pavement width** for all local streets and cul-de-sacs is 28 feet
12.	www.afcee.brooks.af.mil	USA	
13.	www.toronto.ca	Canada	It is recommended that a **pavement width** of 9.4 metres for Royal York
14.	ppc.uiowa.edu	USA	The data show a significant effect of **pavement width** on lane position
15.	www.co.gwinnett.ga.us	USA	**pavement width** shall be at least 24 feet (measured to. back of curb).
16.	www.usace.army.mil	USA	
17.	www.cctrail.org	USA	
18.	www.plannersweb.com	USA	
19.	www.tfhrc.gov	USA	
20.	www.co.honolulu.hi.us	USA	**Pavement width**:. 20 ft.

TABLE X
EXCERPT FROM DOCUMENTS RETRIEVED BY THE QUERY "COOKER PRICE"

Rank	Website	Origin	Comment
1.	householdappliances.kelkoo.co.uk	UK	**Price**: £930 - £949
2.	householdappliances.kelkoo.co.uk	UK	£128.16 - £219.99
3.	shopping.msn.com	USA	$31.99 - $37.75
4.	www.uknetguide.co.uk	UK	Rice Cooker **Price range:** £22 - £22
5.	www.dealtime.co.uk	UK	£1,490 - £1,825
6.	www.business.com	USA	
7.	www.askwhatever.co.uk	UK	
8.	www.currys.co.uk	UK	selling **price** £299.78
9.	www.rangeaway.co.uk	UK	
10.	www.ogormans.co.uk	UK	Mail Order Price £259 FREE UK DELIVERY
11.	www.nextag.com	USA	
12.	www.ogormans.co.uk	UK	
13.	www.electricshop.com	USA	
14.	www.bizrate.com	USA	
15.	www.comet.co.uk	UK	
16.	www.reviewcentre.com	USA	
17.	shopping.msn.com	USA	
18.	www.mysimon.com	USA	
19.	www.ciao.co.uk/	UK	
20.	www.we-sell-it.co.uk	UK	

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf

REFERENCES

[1] Abdelali A, Cowie J, and Soliman H (2005) Language variation as a context for information retrieval. International Workshop on Context-Based Information Retrieval (CIR-05), Paris. July 5th, 2005. CEUR Workshop Proceedings Vol-151, pp. 93-104.

[2] Abdelali, A. (2004) Localization in Modern Standard Arabic. Journal of the American Society for Information Science and Technology (JASIST), Volume 55, Number 1, 2004. pp. 23-28.

[3] Agichtein, E., Brill E., Dumais S., Ragno, R. (2006) Learning user interaction models for predicting web search result preferences, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA

[4] Agirre, E. and Edmonds, P. (2006) Word Sense Disambiguation Algorithms and Applications. Series: Text, Speech and Language Technology, Vol. 33, 2006, ISBN: 978-1-4020-4808-1

[5] Azzopardi L, Girolami M and van Rijsbergen C J (2003) Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures. In the Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval, SIGIR, Toronto, Canada.

[6] Azzopardi L, Girolami M and van Rijsbergen C J (2004) Topic Based Language Models for ad hoc Information Retrieval. In the Proceedings of the International Joint Conference on Neural Networks, Budapest,Hungary.

[7] Cavnar W B and Trenkle M J (1994) N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, pp. 161-175.

[8] Chang, W.W. and Tsai, W.H. (2000) Chinese dialect identification using segmental and prosodic features. Acoustical Society of America Journal. Oct. 2000. Vol.108, pp.1906-1913.

[9] Clarkson P, and Robinson T (1999) Towards improved language model evaluation measures. In: Proc. Eurospeech, p. 2707.

[10] Cowie J, Yevgeny L, and Zacharski R (1999) Language recognition for mono- and multi-lingual documents. Proceedings of the Vextal Conference. Venice 209-214.

[11] Cronen-Townsend, S., Zhou, Y., and Croft, W.B. (2004) A framework for selective query expansion. Poster presentation, in: Proceedings of CIKM'04, pp.236-237.

[12] Dean J, and Henzinger M R (1999) Finding related pages in the World Wide Web. Computer Networks. 31(11-16):1467-79

[13] Dunning T (1994) Statistical identification of language. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.

[14] Gordon M, and Pathak P (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. Information Processing & Management, 35(2), 141-180.

[15] Grefenstette G (1995) Comparing two language identification schemes.Third International Conference on Statistical Analysis of Textual Data. Rome,

[16] Gursky, P., Horvath, T., Novotny, R., Vanekova, V., and Vojtas, P. 2006. UPRE: User Preference Based Search System. In Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence (December 18 - 22, 2006). Web Intelligence. IEEE Computer Society, Washington, DC, 841-844.

[17] House A. S. and Neuburg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. Acoustical Society of America Journal. Vol 62. pp. 708–713.

[18] Ide N, and Macleod C (2001). The American national corpus: A standardized resource of American English. Proceedings of Corpus Linguistics 2001, Lancaster UK.

[19] Kennedy G (1998) An introduction to corpus linguistics. Addison Wesley Longman.

[20] Kohonen T (1997). Self-organizing maps, 2nd Edition (Berlin; New York: Springer).

[21] Lafferty J (1997) The noisy channel model. Class notes to statistical methods in language technologies, Carnegie Mellon University Language Technology Institute, www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/11761-s97/WWW/tex/channel.ps December 22, 2005

[22] Lawrence S, and Giles C L (1998) Searching the World Wide Web. Science, 280: 98-100.

[23] MacQueen J B (1967) Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, Vol. 1. pp.281-297.
[24] Manning C and Schütze H (1999). Foundations of statistical natural language processing. MIT Press. Cambridge, MA.
[25] McNamee P (2004). Language identification: A solved problem suitable for undergraduate instruction. Proceedings of the 20th Annual Consortium for Computing Sciences in Colleges East (CCSCE-04), pp. 94-101.
[26] Moore A (2001) K-means and Hierarchical Clustering - Tutorial Slides. Available at http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html Retrieved on August 29, 2006.
[27] Ponte J M and Croft W B (1998) A language modeling approach to information retrieval system. in Proc. ACM. SIGIR 98, New York, 1998, pp. 275–281.
[28] Purnell, T.; Idsardi, W., and Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. Journal of Language and Social Psychology, Mar 1999; Vol. 18. pp.10-30.
[29] Sethy A, Georgiou P, and Narayanan S (2005). Building topic specific language models from webdata using competitive models. In Proc. of EUROSPEECH, Interspeech, Lisbon, Portugal.
[30] Siatri R (1998) Information seeking in electronic environment: a comparative investigation among computer scientists in British and Greek Universities. Information Research, Volume 4 No. 2.
[31] Spink A (2002). A user centered approach to evaluating human interaction with Web search engines: an exploratory study. Information Processing & Management, 38(3), 410-426.
[32] Torres-Carrasquillo, P. A., Gleason, T. P., and Reynolds, D. A., (2004). Dialect Identification Using Gaussian Mixture Models. In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 297-300, 31 May - 3 June 2004.
[33] W3C (2005) Corpus linguistics. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html.
[34] Wulff S, Gries T S, and Stefanowitsch A (2005) Brutal Brits and argumentative Americans: What collostructional analysis can tell us about lectal variation? Paper presented at the ICLC 2005, Yonsei University.
[35] Yeung K Y, and Ruzzo W L (2001). Principal Component Analysis for clustering gene expression data. Bioinformatics 17, 763–774.

Open Science Index, Computer and Information Engineering Vol:2, No:1, 2008 publications.waset.org/10908.pdf