Multi-Engine Search and Language Translation

Steven J. Simske
Hewlett-Packard Labs
3404 E. Harmony Rd. MS 36
Fort Collins CO 80528 USA
+1 970 898 1359
Simske@hp.com

Igor M. Boyko
Cisco Inc.
Cisco Bldg 8, 3750 Zanker Rd
San Jose CA 95134 USA
+1 650 892 9924
Igboyko@cisco.com

Georgia Koutrika
Hewlett-Packard Labs
1501 Page Mill Rd., MS 1157
Palo Alto CA 94304 USA
+1 650 857 2181
Koutrika@hp.com

ABSTRACT
Two of the most important elements in user interaction with a database are search and language translation. Search is used to access a database system through queries, for which the accuracy and completeness of response are key challenges. Language translation re-purposes content for a different audience, and the accuracy of translated text can be directly evaluated using search output similarity. In this paper, we summarize previously unpublished approaches to improving the quality of both search and translation, with an aim of improving the accuracy of both of these tasks. Specifically, multi-engine and related meta-algorithmic approaches are shown to be promising means of improving the performance of both search and translation. We then describe the vision of how search and translation can be combined to create a more robust overall text mining project.

Categories and Subject Descriptors
G.2.1 [**Mathematics of Computing**]: Discrete Mathematics – *combinatorics*. G.4 [**Mathematics of Computing**]: Mathematical Software. H.3.3 [**Information Systems**]: Information Storage and Retrieval – *information search and retrieval*. I.2.7 [**Computing Methodologies**]: Artificial Intelligence – *natural language processing*.

General Terms
Algorithms, Experimentation, Languages

Keywords
Expert Feedback, Synonym, Meta-Algorithmics, Meta-Analytics, Search, Language Translation

1. INTRODUCTION
Automated search has been a research challenge of high interest to the data mining, knowledge generation and machine intelligence communities for the past several decades. Search queries are comprised of individual textual terms or multiple text terms associated with each other through, for example, a Boolean expression. These queries are often unreliable methods of obtaining the optimal set of documents (or other logical elements) from a corpus. This is due in part to the fact that no default search engine response to a query will provide a customized set of documents matching what a particular user desired in entering the search query. In this paper, we review previous (and unpublished) work on providing the means to (a) extend the search capabilities of a search engine by increasing the likelihood that related documents are found for a particular search query; (b) increase the search efficacy when the user has only a vague idea about what she is trying to find; and (c) provide a means to optimize for several factors how to select documents associated with a query within any corpus. The methods used were part of the body of research to underpin the concepts in a recent book describing meta-algorithmics [1], but were not incorporated into the book.

Meta-algorithmics are a series of approaches to intelligent system design that describe how to combine two or more algorithms or systems into a single system for machine intelligence. Search is a form of machine intelligence associated with filtering; that is, narrowing down a larger body of data into a topic-specific body of data; that is, search output *information*. Language translation is a form of machine intelligence associated with conversion, or *transduction*, of one type of data into another.

The basis of the meta-algorithmic, multi-engine approach to search adopted in this paper was described earlier in patent application [2] which was not exercised. Thus, the research represents previous unpublished research with promising results suggestive of a useful future search research area. The meta-algorithmic approach considered is termed "synonymic search", which allows a single search query to be expanded into a set of queries representing synonymic expressions for the original query. The approach also allows tuning of the amount of synonymic broadening to be applied to the received query for constructing the set of synonymic search queries. Identification of resulting documents responsive to each of the plurality of queries is received, and such received documents are ranked based at least in part on a weighting assigned to each of the plurality of queries.

Language translation, like search, is an important tool for data mining and knowledge generation. In this paper, we present a simple meta-algorithmic approach that combines the output of multiple language translations and uses "expert feedback" in the form of a dictionary in the target language of the translation.

2. SEARCH
Expanding a single search query into a series of related searches is known as query expansion. In this paper, query expansion is incarnated through the use of term synonyms. That is, each term in a query that has one or more synonyms triggers the expansion of the query into a set of parallel queries, each one including only one of this set of synonyms. This process is repeated for every one of N_S terms in the query having one or more synonyms, resulting in a total number of queries, N_Q, given by Equation 1, where S_i is the number of synonyms for term i.

$$N_Q = \prod_{i=1}^{N_S} (S_i + 1) \qquad \textbf{Equation 1}$$

A set of synonymic queries is generated. Many commercially-available, freely-available and proprietary synonym lists exist. WordNet [3], for example, provides the means to generate such a list, and thesaurus options within many word processor engines provide the means to augment the list. Nouns, verbs and adjectives are the common parts of speech used for synonymic queries. In fact, many common articles ("the", "a", and "an"), prepositions ("of", "with", etc.) and conjunctions ("but", "and", and "or", except when the latter two are used in Boolean searching capacity) are ignored altogether in most search engines. Many existing search engines, moreover, separate notions (idioms) consisting of two words into two separate terms, such as in the case of "take off" and "put up" (in which they are treated as "take" and "off" and "put" and "up").

The synonymic search set is typically limited to proximate (and not associated) synonyms in order to keep the number of searches manageable, per Equation 1. Moreover, expressions such as "take off" and "put up" are treated as single candidates for synonyms, resulting in synonyms such as "launch" & "elevate", or "erect" & "construct", rather than synonyms for the individual words in these idioms. Further control over the total number of searches generated is obtained by limiting the number of proximate synonyms, denoted P, to an absolute maximum of, for example, five synonyms (P=5). If there are N terms for which synonyms are found in the original query, there are NP total searches possible. However, to prevent an open-ended number of queries, the total number of queries may be limited to an absolute maximum Q of, for example, 25 queries (most search engines are fast enough nowadays, at several hundredths of a second per query, that this value will typically limit the total search time to < 1 second of searching). The user may also be allowed to limit the total number of searches via a UI device.

Now, if NP > Q, the Q-1 additional searches (the original query is always used) are pruned based on the relative synonymic relationship between each of the terms. An example illustrates this point. Suppose the user types in the query "class list for Stanford". For the term "class", the user will get the following synonyms: set, group, division, grade, rank, category, order (etc.). For the term "list", the user will get the following synonyms: catalog, inventory, register, record, roll, directory (etc.). Already, the number of possible synonymic queries is 56 (that is, 8 x 7), but no more than 25 are allowed (fortunately, "Stanford" is a relatively unique term – although "Stanford University" can be considered a synonym for it, this synonym does not expand the search, and so is ignored). The obvious solution is simply to accept 5 terms for "class" and 5 terms for "list", but this is in general an unsatisfactory solution. Instead, the preferred implementation is to have the synonym database structured such that the synonyms are rated for their "closeness" or "proximity" to the original word. Let us suppose these rates for "class" are 0.9 (set), 0.85 (group), 0.72 (division), 0.65 (grade), 0.51 (rank), 0.42 (category) and 0.23 (order); and for "list" are 0.95 (catalog), 0.9 (inventory), 0.88 (register), 0.85 (record), 0.84 (roll) and 0.46 (directory). The highest 25 combinations are then found by multiplying the synonymic rates together, and so the highest ranking is for "class list Stanford" (1.0), followed by "class catalog Stanford" (0.95), continuing to #24 ("grade catalog Stanford" at 0.6175) and #25 ("division record Stanford" at 0.612).

Note that the "weights" or "proximities" defined above can be further weighted/treated by the "semantics" of the query—i.e. if a query asks, as in the example below, for a "ball sport" then any synonyms of "ball" denoting "dancing" rather than "sports equipment" should be discarded. Such semantic weighting is, in general, quite difficult, and so weighted synonyms such as those demonstrated here help work around this problem. Note that the weights can be defined (a) manually; (b) automatically based on the co-occurrence of such terms in web sites, documents, corpuses, etc. – for instance, reference [4] has a statistical database generated from the British National Corpus, a 100 million word electronic databank sampled from the whole range of present-day English, spoken and written; and (c) automatically based on the order the synonyms occur in a linguistic engine such as WordNet. Almost the same statistical approach can be used for determining the parts of speech (POS) at the front end of query analysis. For example, the word "class" may be a noun, verb or adjective. Using the statistical results from [4], the word "class" is found to be most commonly typed as a noun, and so the appropriate noun synonyms can be used. If, however, a POS analysis of the query indicates that the word "class" is a verb, verb synonyms are found for "class". This is also true of the word "list", which can be both a noun and verb. Since even the best POS engines make mistakes, the user can be allowed to change the POS at the UI level, if available, if they think the engine may have misinterpreted the query.

It is clear from the above that there are numerous approaches to query expansion associated with synonymic proximity along with likely synonymic relevance of the term. After all the search queries have been defined, they are actually run on one or more search engines. On the internet, these search engines can be commercially available ones. On intranets and specific corpuses, they can be whatever search engine the user has available. In this step, all of the queries are provided as input for the search engine and the search engine returns the web sites, documents, etc. that it determines to be best matches. These matches are typically presented in order of relevance, utility, hit frequency, or other reasonable metric, and are presented to the user ranked from 1 to M, where M is the number of "hits" or "matching pages" found.

This approach, "by priority", can use the following types of weighting to combine the search output of multiple engines (note that this is a separate weighting from the query weighting described above): (1) the engines themselves may be weighted by the confidence in the engines; and (2) the order of the results may be weighted, according to their rank in the output set provided by the search engine. It is worth noting, however, that even if a single search engine is used, the synonymic approach effectively provides a multi-engine output. Each is consistent with a meta-algorithmic Weighted Voting pattern [1]. A second means of presenting search output options to users is "by query". This is simple, and has many possible incarnations. For example, each of the original and synonymic searches is presented as a link to the user, and the user can select any of them to find the highest priority sites presented. Another example is to present a tree of the original and synonymic searches [5]. These two approaches have different advantages. The "by priority" approach tends to smooth over biases of a search engine, providing averaging, while the "by query" approach provides quick alternative lists to the user. A preferred motif may be to present the results from the "by

priority" approach with links to the original and synonymic queries in an adjacent column.

An additional presentation mode is possible. In this mode, the overall relevance of all the search results is determined by comparing its keywords to those in the original query. For example, suppose the following two web page descriptions result: (a) a list of people suing Stanford for copyright infringement, and (b) a directory of classes in the Stanford biology program. The first search has "list" at 1.0, "Stanford" at 1.0 and no synonym for class. Its total synonymic weight (using the simplest weighting schema) is thus 2.0. The second search has "directory" for 0.46, "class" (lemma for classes) for 1.0, and "Stanford" for 1.0, for a total weighting of 2.46. Thus, the second search is deemed "more semantically similar" to the original query and is presented higher up in the results. This is the "by semantic weight" approach.

A real example is overviewed here. On one of the major internet search engines, the following query was entered: "ball sport in New Zealand" for which we were trying to find the name of a sport in which you get inside a large plastic double-walled ball and roll down a hill (called "zorbing", a New Zealand invention) and the name for a sport similar to basketball played by women there ("netball"). Both are quite literally ball sports in New Zealand, but they are quite different from the set of top ten results that result for this query in most search engines (almost all are rugby, with basketball or volleyball occasionally making an appearance). The chief synonyms were sphere, globe & orb for ball; game, activity, team game & hobby for sport. The original search "ball sport New Zealand" found chiefly rugby sites, with some hockey and water sports interspersed in the top 10 priority sites. Ditto for "sphere sport New Zealand". When the synonymic search "globe sport New Zealand" was performed, more water sports sites showed up. When "orb sport New Zealand" was queried, zorbing made its first appearance in the high priority list of sites. Water polo appeared when "ball activity New Zealand" was queried; croquet & volleyball when "ball team game New Zealand" was queried; and netball when "ball game New Zealand" was queried. This example illustrates the diversity of returns possible with the use of synonymic query.

3. LANGUAGE TRANSLATION

As a brief introduction to the use of multi-engine translators to increase overall translation accuracy, we used the meta-algorithmic pattern of Expert Feedback [1] where the "expert" was an English language dictionary [4] and the sources to be translated were either in Italian and Russian. Two 500-word documents were hand ground-truthed by the authors and three translation services were deployed (References [6], [7], [8] and [9] for the Italian-English translation).

Table 1. Italian-English Translation

Translator	1	2	3	Combined
Matching %	89.5	91.4	94.1	97.7

Table 2. Russian-English Translation

Translator	4	5	6	Combined
Matching %	80.5	84.9	93.4	96.2

The multi-engine approach for language translation used was straightforward. The words associated with the output of the multiple translations were directly aligned so that the terms could be matched directly for all three translations. Where the translation resulted in non-English words for one or more of the translators, the English word of another translator was used instead. If different English words were identified by the translations, then either the most commonly selected word or else the word provided by the engine with the greatest overall number of successes (English words) was used. This simple multi-engine scheme (Tables 1-2) resulted in reduction of the error rate by 61% (Italian) and 50% (Russian) in comparison to the error rate of the best single engine. Thus, as for search, a multi-engine approach to language translation showed considerable promise.

4. DISCUSSION

Several multi-engine approaches to search and language translation have been demonstrated in this paper. Synonymic and part-of-speech query expansions, in addition to a meta-algorithmic Weighted Voting approach, were shown to provide distinct advantages for customizing search output. Multi-engine alignment and best output acceptance was shown to significantly improve the quality of language translation for two short documents using two distinct languages. Obviously, further quantitative evaluation of the approaches outlined herein will be a useful next set of experiments. This paper only highlights the large set of possibilities in this space. In the future, validation of the techniques with IR datasets from TREC (queries and qrel ground truth) will be performed. The approaches outlined here will also be compared to relevant similar approaches, including query expansion, exploitation of synonyms and cross-language IR. Finally, it should be noted that there is a logical link between these two fields of text data filtering and transduction. Namely, the accuracy of the language translation approach can be directly gauged by comparing the search results on the un-translated and subsequently translated corpora. If the translation is accurate, then the documents should respond very similarly to un-translated and translated *queries* against the corpora. This type of *functional* testing of un-translated and translated corpora also warrants further, quantitative investigation. This will be a focus of future research for our team and, hopefully, others.

5. REFERENCES

[1] Simske, S.J. 2013. *Meta-algorithmics: patterns for robust, low cost, high quality systems*. Wiley & Sons, Hoboken, NJ, USA, 386 pages.

[2] Simske, S.J. and Boyko, I. 2002. System and method for management of synonymic searching. US Patent Application 10/256,674 (20040064447).

[3] Word Net, http://www.cogsci.princeton.edu/~wn/.

[4] British National Corpus, http://www.natcorp.ox.ac.uk/.

[5] Vivisimo, http://www.vivisimo.com.

[6] http://inews.tecnet.it/show.asp?f=articoli/2002/04/IN0204_Focus_Kids.htm.

[7] Free Translation, http://www.freetranslation.com/.

[8] LinguaTec, http://www.linguatec.net/online/.

[9] WorldLingo, http://www.worldlingo.com/wl/Translate.