Utilizing Sentence Function Tagging to Retrieve Related Text Excerpts

Tim Musgrove & Robin Walsh

TextDigger, Inc.
305 Vineyard Town Center #375
Morgan Hill, CA 95037
{tmusgrove,rwalsh}@textdigger.com

Abstract

A method is described for utilizing a fast, rule-based thematic tagger on excerpts of texts within a SERP (Search Engine Results Page) as a context model for finding snippets from deeper in the SERP that are relevant to those found higher up. This method enables a "more like this" functionality which is found to have accuracy materially better that that accomplished merely by keyword matching, without introducing too much latency for the user.

Background

Whenever users perform keyword searches over large document collections, they usually are left wanting more than what their initial search results contain. A common method of assisting users in this predicament is to provide a link to additional related documents, in connection with one of the particular "hits" on the search engine results page (SERP). Sometimes this option is manifested in the user interface as a "More like this" hyperlink, placed next to each numbered hit in the SERP. Typically such a link will bring up documents that have surface characteristics in common with the reference document from which it originated, such as the same or similar content words in their title, together with the same or similar hit keywords from the user's query [e.g. Alani 2001]. If a topic category is known, the functionality may involve a match of the category as well.

We believe that many such systems suffer from an insufficient model of semantic context. After explaining what these problems are, we then delineate the requirements for a better method, and describe our early attempts to prototype a system that satisfies such requirements.

There are problems with recommending documents in an overlapping domain or category, even when they share many similar keywords to a user's query. An automobile site concerning the history of automobiles, and another concerning their repair, may both be members of the same general domain of "automobiles" and may both share common keywords such as "Ford" and "performance" and "engine," despite one being completely irrelevant to users' interests in the other sub-domain. How do we prevent our system from recommending one sort of document to a user looking for the other?

What seems to be called for is more robust modeling of the text excerpt which had drawn the user's interest in a hit list. For when the user clicks on "More Like This" in a SERP, they are not clicking on just that link: they are clicking in reference to an excerpt of some text, and in that excerpt, sparse though it may be, are words that form a crucial semantic context. It stands to reason that we could analyze these text excerpts, as a means of guiding or filtering the recommendation of other relevant texts for the user. Usually the only "model" of these contexts is a vector of keywords contained in them. While being helpful in many cases, such a limited model under-utilizes the available context of the user's action of clicking "More like this." It tends to produce results exhibiting much the same limitations and deficiencies as the original keyword search did, as it is based on much the same methodology.

In order to really behave "as advertised," a "More Like This" link should not only match the query words and belong to an overlapping domain, but also exhibit some thematic elements of the text excerpt corresponding to the user's action, i.e. the context in which the user clicked "more like this." To do this, applying a full-scale parser and extracting a rich set of semantic roles would be a fairly obvious approach. For example, one could then match excerpts in which the same keyword is a subject noun, as indicated by the parser (rather than, say, as an object of a preposition); or one could match hits that contained the keyword in the <Actor> role rather than in the <Instrument> role, as indicated by a semantic role analyzer. These methods would surely be promising.

However, there are several pragmatic considerations that must be brought to bear on such proposals. First, the application of a robust parser (e.g. Charniak 2000, Collins

---

Some of the technology described herein is covered by patents pending for TextDigger, Inc.

1996, or Lin 1998) and/or a good semantic role analyzer (e.g. Gildea 2002) is prohibitive in terms of real-time latency and computing cost, for the type of application in which we are most interested (i.e., a broad, large scale search for average users).

Another option would be to establish an ontologically faceted collection ahead of time, such as Collex [Novisky 2005], but this requires a substantial editorial investment and therefore is not applicable to our general search application. Instead of these methods, we need to apply some lightweight method of extracting useful themes from un-annotated texts.

Secondly, there are constraints upon accessing the full text of documents referenced in the SERP. The full texts are for all practical purposes not available in the use-case scenario that we are examining, for it is very expensive to retrieve full documents for analysis in real-time.

On the other hand, it is fairly cheap in terms of computing resources and time, to retrieve a SERP containing up to two hundred candidate hits, each manifested as a text excerpt containing the keywords which the user searched for. Knowing that the typical user will look through only the first ten or twenty of these excerpts, we have perhaps 180 additional ones which could hold some real gems, as it were, for a "More Like This" option. It is in these excerpts that we must hope to very quickly extract the needed thematic elements to provide a smarter "More like this."

Explanation of Method

Our primary approach is to place a wrapper around the user's chosen keyword search engine, and (1) initially retrieve the first 200 hits, (2) after taking the top 20 hits as reference excerpts, analyze the remaining 180 to find which are "more like" the top twenty, and then (3) present the top twenty alone to the user, with the "more like this" links having just been inserted. This all needs to happen very fast in real time, and without the requirement for any previous offline manual tagging.

Besides failing to meet the speed requirement of this application, many parsers and semantic role analyzers expect full sentences that are largely grammatical. By contrast, a typical SERP bears only short excerpts of text, really just snippets of sentences. We need to glean what we can from the excerpts in the SERP itself.

To develop a prototype of such a system, we obtained a lightweight sentence function tagger, called "SIFT" (for Sentence Intentional Function Tagger, licensed from MTE, LLC.), and implemented it in correspondence with an Internet keyword search. The tagger executes a rule-based partial parse to produce a variety of speech-act and assertion-type tags on sentences, such as "interrogative," "causal connection," "supporting example", "description", "definition", "warning", "emphasis," "inference," and so on. Examples of such rules are shown below.

Example SIFT rules (where 'A' = assertion, 'P'=phrase, 'T' =term):

Tag	Trigger	Object(s)
Inference	"Since" A1, A2.	A1,A2
Warning	"Watch out for"… P1…	P1
Definition	T1 ["consists of | is comprised by"] Noun-List.	T1

Since these are domain-independent functional tags and are not strongly content based, they provided a computationally light method of identifying, across a broad range of topics, several good modeling characteristics for the excerpt sentence (or snippet thereof) in which the "hit words" were found. This SIFT approach was comparatively faster than applying full parsers or semantic role analyzers because it does not analyze more than a small set of semantic elements from a rather shallow parse.

Furthermore, SIFT's partial parse works fairly well on mere snippets of sentences, which are what we see in many SERP's (though, for perspicuity, we will show snippets as full sentences in the examples discussed herein). This means we can afford to analyze every text excerpt on the SERP in real-time.

Since the definition of our problem-space includes the presumption of topic-matching, we selected a particular domain ("sports injuries") and started applying the SIFT tagger to excerpts outputted in the SERP by a keyword search on the Web in this topic area. We soon found, in implementing SIFT for our special purpose here, that some of the function tags it produced were more useful than others. For example, <Quantification> and <Description>, while intellectually interesting, were not particularly helpful for our application. They are too easy to satisfy, and snippets bearing no special relevance could easily exhibit these features. On the other hand, functions such as <Causal connection>, <Avoidance goal>, <Diminution> and <Warning> were very helpful. So we manually flagged the functions we were interested in, ensuring that the others would be ignored.

Next, we found that grouping certain similar functions together was helpful, for example, <Contributory cause>, <Logical connection>, and <Causal connection>. Such a grouping was treated disjunctively in order to widen the pool of eligible sentences from foreign documents. This widening of the pool was critical in obtaining a respectable coverage rate. For example, for a given sentence having a <causal connection> on a particular keyword, there might be only a single snippet exhibiting the same features. However, when including sentences that also had a <logical connection> or <contributing cause> connected

with that keyword, the pool might increase from one candidate sentence, to several. This type of example repeated itself many times.

However, after widening the pool in that respect, we found there was a need to narrow the pools in a different respect. Namely, we found that for some functions, unless sentences had (semantically) similar elements addressed by their common function, then they were not as likely to be relevant to one another, e.g. while there was obvious relevance between a sentence asserting a causal connection pertaining to "pain" and another sentence establishing a causal connection pertaining to "hurt", neither of these were relevant to a sentence ascribing causality to something else unrelated to hurt or pain, for instance, to rule changes in a sport. So we established in some cases that having the same function in common was not enough, but that the function should apply to either a shared topic word (literally, a word found in the topic name or one of its aliases, after stemming), or a query word, or else another word that occurred in both snippets (or that has a synonym or hyponym in each snippet).

Therefore we flagged some functions as requiring "strong matching". For these functions, we hoped the snippets would both have the same function connected with the query word or one of the topic words, but if not, we called WordNet to check if there was close semantic distance[1] (e.g. synonymy or hyponymy) between words connected in the two corresponding snippets to the common function. So for example, two sentences with the keyword "head" exhibiting the <Emphasis> function with elements "severe" and "serious" respectively were deemed a strong match, given that WordNet lists "severe" as a synonym for "serious." This is opposed to other words that could have manifested the <Emphasis> function tag such as "especially," which is not a synonym for "serious" (nor is it a synonym for "seriously") and therefore would not count as a strong match.

Here is an example of a context model, for one particular search result:

SERP hit excerpt (query terms in bold):
…very serious **concussions** or neck injuries can result when proper tackling is not… other **football** injuries can be prevented by utilizing the right equipment and teaching proper techniques…. coaches are advised to drill for at lest two weeks before the first scrimmage

Context model:
Topic terms: sports,injuries
Query terms: concussions, football
Tag: Causation; **trigger: "**can result when"; **objects: "**proper tackling is not" / "neck injuries"
Tag: Emphasis; **trigger**: "very serious"; **object**: "concussions"
Tag: Avoidance; **trigger**: "can be prevented"; **object**: "other football injuries"
Tag: Warning; **trigger**: "advised to"; **object**: "drill"
Tag: Instrumentality; **trigger**: "utilizing"; **object**: "the right equipment"

Note that the combination of the aforementioned topic terms, query terms, and tags (with their objects), comprise the complete context model for the SERP hit in question.

This example illustrates that our context model shares some elements with other available models. It begins with some simple keywords (those of the topic name and the user query), which are no more than the "baseline" context model. Next it has some commonalities with speech acts theory, e.g. the "warning" label. Finally, it overlaps with semantic role models in the case of the "instrumentality" tag. Many semantic roles are not included in the system, so the overlap is not complete.

Given that the context model has partial overlap with a plurality of other models, it can be deemed eclectic. Nonetheless, all of its tags have one important thing in common: they are obtainable quickly from simple rules without need for a complete grammatical parse of the text or exhaustive semantics analysis.

In our experiments, after finding numerous function-matches according to the procedure above, we were ready to make recommendations for "more like this". For a given reference excerpt, those candidate excerpts (i.e., lower hits in the SERP) having at least one function match were selected for recommendation, and in cases of multiple matching candidates, those with the most function matches were listed first. In the case of two candidates bearing the same number of function matches to the reference excerpt, the excerpt that had ranked higher on the original SERP was deemed the winner of the tie.

Below are two examples that resulted after enforcing all of the rules described above (query words highlighted in bold):

Example 1:
Reference sentence: "Numerous methods to reduce knee **swelling** have been advocated, including elevation of the limb, ice therapy, Quadriceps muscle exercises, massage and electrotherapy treatment."

Function-matched sentence: "Once the doctor knows the full extent of your injury, he or she usually will start with conservative treatment techniques such as rest and ice to help decrease **swelling**."

Explanation of function-match: <Diminution> tag applied to reference sentence for having "reduce" connected to "knee swelling"; same tag applied to matched sentence for having "decrease" applied to

[1] For various ways of calculating semantic distance see [Alani 2000].

swelling; "reduce" a hyponym of "decrease" in WordNet.

A candidate sentence that failed to function-match: "Some otolaryngologist-head and neck specialists set fractured bones right away before **swelling** develops, while others prefer to wait until the **swelling** is gone."

Example 2:
Reference sentence: "The damage caused after one **concussion** is often reversible after an appropriate recovery time, but if a second injury is sustained before then, the damage can be devastating."

Function-matched sentence: "Every head injury should be taken seriously and it is important to understand that the damage done by multiple **concussions** can be cumulative."

Function-match explanation: <Repetition> tag applied to reference sentence for having "a second injury" where "injury" is a topic-word; matched sentence received same tag for having "multiple" preceding query-word "concussions".

A candidate sentence that failed to function-match: "A **concussion** is a violent jarring or shock to the head that causes a temporary jolt to the brain."

It is important to note, in the above examples, how the failed sentences are less relevant, even though they belong to the topic of "Sports Injuries" and also contain the query word. Such cases were plentiful. There seems to be a variety of ways in which a sentence can be irrelevant despite belonging to the right topic and containing the same query word(s), whereas it is difficult to find a sentence that can instantiate a function-match meeting all our criteria, without being relevant.

Testing

After being convinced that we had tailored our implementation of the SIFT tagger as best we could for this application, we showed a sample set of function-matched sentences to a number of users. We presented these as triples of sentences, each triple consisting of a reference sentence followed by a pair of recommended sentences from topic-related documents, wherein one candidate sentence had been selected by our function-matching method, and the other selected merely by keyword match, i.e. the sentence that shared the most content words (other than query words) in common with the reference sentence.

We asked respondents to select whether one sentence or the other was more "relevant" to the reference sentence. Respondents also had the option of indicating that both sentences were equally relevant, or that neither sentence was relevant to the reference sentence. Out of our sample of possible recommended sentence pairs we received the following results:

Function-matching sentence deemed more relevant	Keyword-matched sentence deemed more relevant:	Both sentences deemed equally relevant:	Neither sentence deemed irrelevant:
57%	10%	27%	7%

There were two interesting observations from these results. First, in a very large portion of cases (43%), there was no benefit to the function-matching, because it was deemed either to be irrelevant, less relevant, or no more relevant than a simple keyword matched sentence. However, in cases where one sentence was deemed better, the function-matched statement almost always was the victor, being favored 57% of the time, as opposed to only 10% of the time for keyword-matched sentences. This leaves us with the overall conclusion that the function-matching seldom injures relevance, and around half the time, it helps.

Next we wanted to see how the system performed in terms of overall accuracy. We knew from the foregoing test that the precision was around 83%, but for recall, we examined over 300 sample sentences against a smaller set of reference sentences. After human-tagging all sample sentences deemed relevant to each one of the reference sentences, we found that the function-matching system had a 68% recall rate. By contrast, keyword matching alone produced 88% recall but only 37% precision on the same sample. Our f-measures, i.e. the harmonic mean of precision and recall, *a la* [Van Rijsbergen, 1979], are shown here:

F-measure for function matching sentences:	F-measure for keyword matched sentences:
0.75	0.52

Conclusion

In an environment requiring speedy processing of SERP's, we found that given a particular item in the hit list, it is possible to recommend sentence-level relevant hits from related documents further down in the list, on the basis of a lightweight sentence-function tagging engine, with an accuracy rate materially greater than that of merely matching on keywords (albeit with somewhat lower recall). The result suggests that contextually relevant hits from related documents could be automatically suggested on the basis of sentence-level clues, provided that there is a sufficient overlap in their sentential functions, and a large enough pool of excerpts for the system to examine.

References

Alani, H., Jones, C. and Tudhope, D. (2000) "Associative and Spatial Relationships in Thesaurus-based Retrieval". *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL2000),* edited by J. Borbinha and T. Baker, Lecture Notes in Computer Science (Berlin: Springer), pp. 45-58.

Alani, H., Jones, C. and Tudhope, D. (2001) "Augmenting Thesaurus Relationships: Possibilities for Retrieval". *Journal of Digital Information*, Volume 1 Issue 8, Article No. 41.

Charniak, E. (2000) "A Maximum-Entropy-Inspired Parser". in *Proceedings of NAACL-2000*, pp. 132-139.

Collins, M. (1996) "A New Statistical Parser Based on Bigram Lexical Dependencies". *Prcoeedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA, p. 184-191.

Lin, D. (1998) "Dependency-based Evaluation of MINIPAR". In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May, 1998.

Fellbaum, C., ed. (1998) *WordNet: An Electronic Lexical Database*, MIT Press. ISBN 0-262-06197-X

Gildea, D. and Jurafsky, D (2002) "Automatic labeling of semantic roles". *Computational Linguistics* 28(3):245-288.

Nowviskie, B. (2005) "COLLEX: semantic collections & exhibits for the remixable web," a preprint, November 2005. http://www.nines.org/about/Nowviskie-Collex.pdf

van Rijsbergen, C.J. (1979) *Information Retrieval*, 2nd Edition. Available online at: http://citeseer.ist.psu.edu