Internet search mechanisms and distortions of the semantic space: the scientific challenges facing the "Googles"

A. Linhares & C. W. Afonso
Getulio Vargas Foundation, Brazil

Abstract

Ever since the launch of Altavista, internet search engines have become a multi-billion dollar industry, with fierce competition between Google and the three major competitors. One of the challenges involved is to rank search results in a way that places the most meaningful results at the top. In order to do this, the algorithms involved must try to grasp the actual meaning, the semantics, embedded in a search query. In this paper we discuss a problem we call "distortions of semantic space". Distortions of semantic space occur regularly in people's texts, writing styles, labeling of images, etc. We present a number of examples of distortions of semantic space, and analyze the problem. We also comment on new computational architectures that have tried to handle this problem, albeit the state of the art still remains far from the needed challenge.
Keywords: search mechanisms, distortions of the semantic space, Google, literal search, new contents of the internet, semantic web.

1 Introduction

Google defines its mission as "to organize the world's information." Since its launch, in 1998, it has reached enormous financial and marketing success, given its superior ranking and indexing technology of data in the Internet. It is now possible to carry searches in 100 different languages with Google, and in 2005, the company reached the mark of *a billion searches per day* (Friedman [1]). To sustain this leading strategic position, however, the company faces enormous scientific obstacles so that, as the types of information available on the web change, new technologies may be able to organize them in an agile form for all to access.

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070171

1.1 Organizing the world's information

One of the greatest landmarks in the evolution of the Internet was the appearance of search mechanisms such as Google, which quickly succeeded Altavista in market leadership. The gigantic amount of information available on the web was, previously, of difficult access; as sites such as Yahoo! or Internet Yellow Pages (today only of historical value) tried to organize such data using a directory structure, cataloguing each page and site according to the interpretation of their employees. Two problems emerge with this approach:

(i) The interpretation of the employee who initially catalogued the page could be different from the interpretation of the user; suppose an employee categorized eBay, the giant auction website, in a /shopping/auctions directory structure. Imagine now that a specific user were searching for a "place to find people who collect stamps". eBay obviously is such a place; however, classification through a directory structure cannot lead all its potential user base to it.

(ii) the scalability of the model, as the number of pages available grew from a few hundreds, to thousands, then millions, to today's billions. It is not economically viable to pay large amounts of people to catalogue billions of pages, and, even if it were, that would be a *Sisyphus* task, as these pages are in constant content change.

As we will see below, these factors enabled Google to conquer a significant part of the added value in organizing the internet's information.

1.2 Strategic sustainability: the best results in the top

Two questions are crucial to understand the success of Google and the sustainability of its strategy. (i) Why is Google the leader of the search market? (ii) What supports Google in that leadership position?

Why does Google lead the market of searches? The first-mover advantage assumption is, in this case, simply wrong, as Google had at least 7 previous mechanisms in the brief history of the WWW:

(i)WWW Wanderer
(ii)WWW Worm
(iii)Webcrawler
(iv)Lycos
(v)Infoseek
(vi)Excite
(vii) Altavista

These two initial engines considered only page headers, and not the pages' main content. Altavista, launched by the research department of Digital Corporation as demonstration of the power of its 64-bits "alpha" processor, was the first engine to consider the entire content of all pages in the Internet – which guaranteed the leadership of Altavista until the launch of Google. Unfortunately for Digital Corp., Google possessed basic characteristics that would enable it to quickly surpass Altavista. *These characteristics are the target of our work, and will be dealt with in section 2.* Today the search mechanisms that divide market share are:

(i) Google
(ii) Ask (formerly Ask Jeeves)
(iii) MSN Search (Microsoft)
(iv) Yahoo! Search

The second question involved is: what supports the company in this position of market leadership?

1.3 The value of the service

What is the value of Google services? One of the forms to evaluate the company is to verify its financial market value.

Recent fluctuations of Google stocks are presented below in Figure 1. After reaching a maximum above USS470, the value had fallen, in May of 2006, to USS370.

Figure 1: Google's market capitalization based on its stock value surpasses 100 Billion dollars. The company, who approximately possesses revenues of six billion dollars, is evaluated as more valuable than Petrobras, whose revenues surpass 50 billion dollars (the reader should also consider that the Petrobras stock value has grown considerably due to increases in international prices of crude oil).

2 "Intent drives search": from psychology to new mechanisms of search

"Search is a problem 5% solved", says Udi Manber, the CEO of the search mechanism A9 from Amazon.com (Batelle [3]). In this section we explore the nature of the search problem. Not all searches are for a determined topic of a subject (Batelle [3]). Approximately 15% of the searches look for a good set of links, in contrast to a good document. Approximately 25% of the searches are navigational, that is, for a specific website that the user already had in

mind. About 36% of the searches are made with sights to a transaction, either commercial, or information on tracking of a package, etc. Approximately 12% of the volume of searches is referring to sex. Given this variety of initial intentions, we can start to visualize why the problem is only 5% solved (Battelle [3]). Engines still have to consider that the common user generally type only one or two terms in a search.

Let us assume, for example, a search for the word "jaguar". Which type of pages must appear as the first ones? Consult the word Jaguar on Ask.com, and it will guide the user for a definition of the *original intention*: "animal jaguar"?; "Car"?; etc. Our mind is extremely fast in processing ambiguous information and reinterpreting them within the context (Hofstadter [4]; Linhares [5]). As example, the ambiguous phrase "prostitutes appeal to pope" is mentally reorganized after an initial interpretation, but engines lack this reinterpretation capability. A search for "biography Abraham Lincoln" does not mean that a desire for all biography pages mentioning "Abraham Lincoln".

Symbols and semantics: What is it really desired from one or two requested words? Given one or two symbols, which is the meaning that you looked for? If one wants to understand what Manber had in mind when he said that the problem is 5% solved, we can observe a quote from Batelle [3], mentioning the problem of understanding original intention:

> *But how might we get there? For search to cross into intelligence, it must understand a request--the way you, as a reader, understand this sequence. [...] My problem is understanding something. That can only happen if search engines understand what a person is really looking for, and then guide them towards understanding that thing, much as experts do when mentoring a student.*

This problem seems simple, yet, it is daunting. Consider, for instance, the question "What is similarity?", as it applies to text documents such as those indexed by search engines. If Google found two documents with thousands of words in exact sequence but a mere comma of difference between them, should the engine classify such pages as similar? It seems obvious, for there is no reason the algorithm might dismiss a mere comma to make any significant change in what concerns the content of the documents. Then again, consider it from a human's eyes.

This is an intriguing example, from a story reported on the New York Times (Ian Austen, The Comma That Costs 1 Million Dollars [Canadian], October 25, 2006):

> *The Comma That Costs 1 Million Dollars (Canadian)*
>
> *OTTAWA, Oct. 24 — If there is a moral to the story about a contract dispute between Canadian companies, this is it: Pay attention in grammar class.*
> *The dispute between Rogers Communications of Toronto, Canada's largest cable television provider, and a telephone company in Atlantic Canada, Bell*

Aliant, is over the phone company's attempt to cancel a contract governing Rogers' use of telephone poles. But the argument turns on a single comma in the 14-page contract. The answer is worth 1 million Canadian dollars ($888,000).

Citing the "rules of punctuation," Canada's telecommunications regulator recently ruled that the comma allowed Bell Aliant to end its five-year agreement with Rogers at any time with notice.

Rogers argues that pole contracts run for five years and automatically renew for another five years, unless a telephone company cancels the agreement before the start of the final 12 months.

The dispute is over this sentence: "This agreement shall be effective from the date it is made and shall continue in force for a period of five (5) years from the date it is made, and thereafter for successive five (5) year terms, unless and until terminated by one year prior notice in writing by either party."

Consider that last comma. How long should the contract last? Without the comma, it's pretty clear, right? It must last at least a full 5 years. It is beyond the point whether the lawyers actually intended this, but the comma, however, distorts meaning in a profound way. This distortion of meaning brought by the slightest of cues is a significant cognitive phenomena, for it happens, many times, subconsciously in a human's information-processing, with no need for any conscious thought.

Let us now get back to Google's way of looking at things. There are two 14-page documents, one has a single comma that the other lacks. Should Google classify them as "similar"? It seems clearly obvious that it must be the case: to Google's eyes, these are 99,9999% similar. After all, under what circumstances should the algorithms in a search engine perceive the semantic dangers that lie within a single comma, given thousands and thousands and thousands of exactly-matching-words-and-paragraphs documents?

3 Focusing the problem

In this section we discuss the nature of some problems regarding search mechanisms. We initially consider the problem of literal search, and later, the problem of search for multimedia content of dauntingly difficult indexation.

3.1 Literal search

In 1957, a thought by J.R. Firth launched an idea well used in the study of linguistics, which later would influence the mechanisms of literal search: "*You shall know a word by the company it keeps*".

Behind this phrase is the idea of *correlations between words* that help understand the meaning inherent to each word. Words with similar meanings would tend to appear in a great number of texts, and, therefore, its meaning could be extracted from the analysis of the relations between words. In fact, this was the idea used in search mechanisms. This seems to be a simple mechanism for extracting intent, yet, we claim that the mechanisms of literal search face four basic problems:

(i) Deformations of the semantic space - similar words are considered next in the semantic space. Through the process of analogies we perceive an object as pertaining to another class of objects. An mp3 player, of Apple, iPod, can be seen as "walkman", but also it can be seen as "a printer", or "ferrari", or a "Trojan horse" (Afonso and Linhares, [6]). Another example given by French [7]: the word "hammer" is next in meaning to saw, nails and other construction materials, but one is capable of attributing different meanings to the same objects. The hammer can as a paperweight, losing its initial function (and starting to become related with different objects) in semantic space. Linhares and Brum [8] have shown that this effect arises in chess players' strategic thinking.

(ii) The mechanisms of literal search do not detect the occurrence of abstract structures - through the process of analogies we compare different things: an iPod "is a Ferrari of mp3 players"; "Google is the new Microsoft", etc.

(iii) The mechanisms of literal search do not know the words in the same way we do - we know the words through experience and contact with the world, which makes them assume multiple meanings and connotations to us; but not to search engines.

(iv) They consider that words are atomic entities- for human beings words are not atomic; therefore syllables can assume distinct functions, which complete the meaning of the word.

It is due to these four basic problems that the thesis of correlation between words helping to understand the meaning inherent to each word may be discarded. We can see that in some examples: when we ask the system "a good name for Father", the word "John" appears, obviously. But the word "Mary" also appears. More interesting: when we ask " a good name for the prime minister of Israel", the words "Sharon", "Isaac", "Rabin", appear in the top. But also appears "Arafat". Why? Because the searches are made based on correlations, and Arafat obviously is correlated to "prime minister of Israel" in millions of texts in the web.

The systems of literal search are blind for certain connections that we make easily. Hofstadter [4] discovered that our mind is only capable of understanding things because it perceives, impulsively, subconsciously, abstract roles for words and things; therefore we use so many analogies. When we ask the system to classify "how much you perceive lawyers as":

(i) telephones
(ii) sharks
(iii) blood suckers
(iv) vampires
(v) rocks

The system says that lawyers are more "telephones" than "vampires" or "bloodsuckers", when most people respond otherwise. Why does the system make such erroneous mistakes? Because it is blind to the abstract roles that we see lawyers portraying in our society. The system is incapable of making analogies that we make immediately. What we, human beings, see, when we understand what we see, are abstract roles that allow us to make analogies [4, 5, 9]. Let us see some examples.

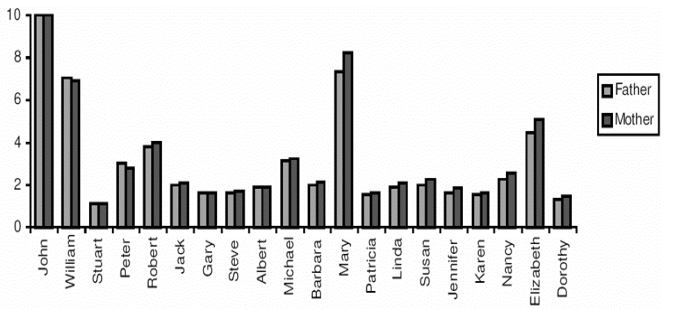

Figure 2: What is a good name for a "Father"? What is a good name for a "mother"? As these words tend to appear in similar contexts (example: "the mother of Jack"), the results are very similar for both sexes.

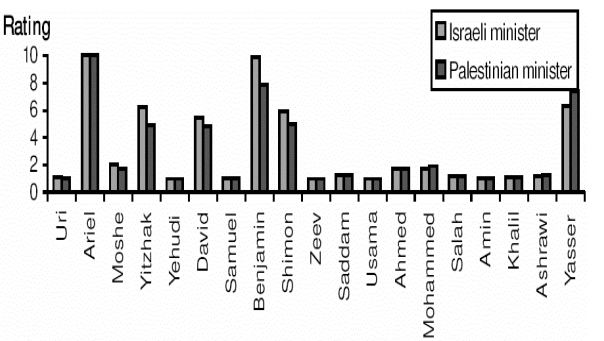

Figure 3: Which is a good name for prime minister of Israel? Which is a good name for prime minister of Palestine? As in the example above, the proper names are correlated with both sides, so that Saddam Hussein seems a good name for prime minister of Israel ("prime minister of Israel threatened Saddam Hussein...").

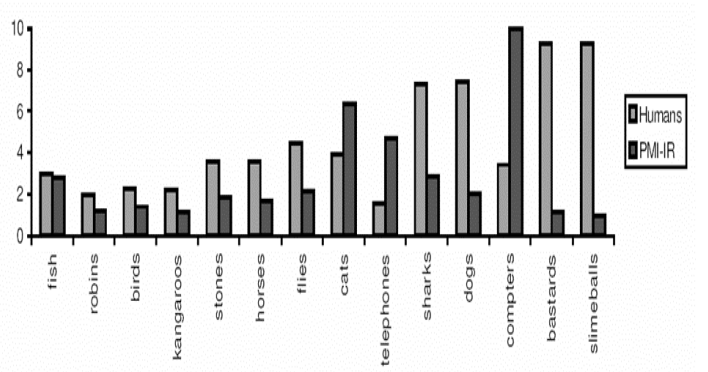

Figure 4: When we asked the system to "rate lawyers as: horses, fish, telephones, stones, sharks, cats, flies, birds, slime balls, kangaroos, robins, dogs, and bastards", the results are the opposite of what humans think.

The skeptical reader could argue: "does this type of anomaly occurs in practical situations? Could a system such as Google really offer this type of results?" Let us see one example of the following search: "Israeli prime minister name" (carried through in May 30th of 2006). Between 'top ten hits', we can find:

> **BBC NEWS | Middle East |** Hamas 'names its prime minister'] Israel says it will not deal with a Hamas government unless it renounces violence ... We have decided to nominate brother Ismail Haniya as prime minister ...**news.bbc.co.uk/2/hi/middle_east/4721456.stm - 41k**

It is indeed the case that search engines are 'fooled', and point out exactly the enemies of those intended in the search query! Since semantics changes subtly, it is incredibly hard for current architectures to perceive subtle distortions of meaning. Consider, for instance, irony: what kind of search could bring ironic pages (without explicit mention of the word irony or related terms)?

Despite such a profound difficulty, these problems have began to be approached in the last decade. These slight distortions of semantic space have been dealt with "Fluid concepts" architectures. Let us look at an example below.

4 Fluid computational architectures

A number of computational architectures employing *fluid concepts*, which can account for these distortions of semantic space, have been devised recently [4, 9, 12]. Our own investigation into this problem has shown that the same effect of distortion of semantic space arises in chess positions. Positions which have distinct superficial features can be perceived as highly similar at a strategic level,

while positions that share many superficial features can still be seen as extremely different in strategic terms. Figure 5 presents, for example, two positions that differ superficially while still maintaining enormous strategic similarity. In the left position, white moves the rook to g8 check, black captures with f8—g8, then white checkmates with knight to f7. In the right position, white moves the knight to a6, placing a (double) check. The black king must then escape to a8, to which white responds with queen to B8 check. Once again, black counters with the rook capture and white checkmates with knight (back) to c7. The positions are similar at a semantic level, while completely different at a superficial level.

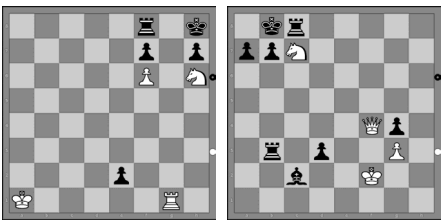

Figure 5: White to move and win. See the text for the solutions.

The current implementation of the computational architecture follows the philosophical foundations of the architectures posed in [4]. The key idea of this project (figure 6) is to build a computational architecture to model a player's high-level, abstract perception of a given chess position, during the first fleeting seconds it is perceived. The architecture displays (i) a high degree of 'entropy', constructing and destroying structures as new perceived roles receive higher priorities; (ii) a high degree of parallelism, with both bottom-up and top-down processes (termed impulses below) running concurrently; (iii) concurrent processing distributed over multiple levels (pieces perceived, distance relations, chess relations, abstract roles, etc); (iv) and "vagueness" brought by continuous degrees of intensity of abstract roles.

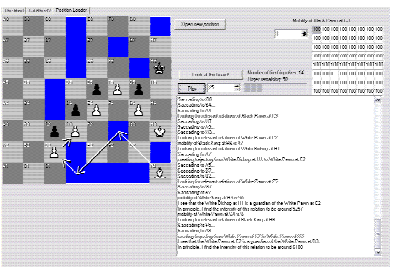

Figure 6: The project at an early stage of processing a position, even before many squares have been 'saccaded to'.

References

[1] Friedman, T. *The world is flat*. New York: Farrar, Straus and Giroux, 2005.
[2] The Economist. Is Google the new Microsoft?, *Leaders column, Information technology*, May 11th, 2006.
[3] Battelle, J. *The Search: how Google and its rivals rewrote the rules of business and transformed our culture***.** New York: Penguin Books, 2005.
[4] Hofstadter, D. (1995) *Fluid Concepts and Creative Analogies*. New York: Basic Books, 1995.
[5] French, R. M. and Labiouse, C. (2001). Why co-occurrence information alone is not sufficient to answer sub cognitive questions. *Journal of Theoretical and Experimental Artificial Intelligence*, 13(4), 419-429.
[6] Afonso, C. W. and Linhares. Analogias no processo decisório. *Working pape,* 2007.
[7] French, R. M. & Labiouse, C. Four Problems with Extracting Human Semantics from Large Text Corpora*. Proceedings of the 24th Conference of the Cognitive Science Society,* 2002.
[8] Linhares, A. and Brum, P. Understanding our understanding of strategic scenarios: what role do chunks play? Accepted for publication, *Cognitive Science*, 2007.
[9] Linhares, A. An active symbols theory of chess intuition. *Minds and Machines* 15, pp. 131—181, 2005.
[10] Linhares, A. A Glimpse at the Metaphysics of Bongard Problems, *Artificial Intelligence* 121, pp. 251–270, 2000.
[11] Handler, J., Berners – Lee, T. & Lassia, O. The semantic web. *Scientific American*, 284(5):34–44, May, 2001.
[12] Linhares, A., New ideas for computer chess, submitted for publication, *IBM Systems Journal,* 2007.