A Communicative Approach to Web Communication: the Pragmatic Behavior of Internet Search Engines

Lorenzo Cantoni

NewMinE Lab

University of Lugano

via Buffi, 13 – CH 6900

Lugano

+41 58 6664720

lorenzo.cantoni@lu.unisi.ch

Marco Faré

webatelier.net

University of Lugano

via Buffi, 13 – CH 6900

Lugano

+41 58 6664788

marco.fare@lu.unisi.ch

Stefano Tardini

eLab – eLearning Lab

University of Lugano

via Buffi, 13 – CH 6900

Lugano

+41 58 6664527

stefano.tardini@lu.unisi.ch

ABSTRACT

In this paper, websites are not approached as being just technological artefacts – which they also are, indeed – but from the point of view of communication, which is (one of) their structural purpose(s). In this perspective, the Website Communication Model (WCM) provides a model that highlights five main areas of interest when dealing with websites: the areas of contents and services offered through the website, of the tools for accessing them, of the people who publish the website, of those who access and use it, and of the "ecological" context which the website is part of.

The need for such an approach to electronic communication is well represented by the behavior of internet search engines, which strongly rely on the 'pragmatic' aspects of web communication. In fact, when performing the activities of collecting web pages, indexing them into their databases, and responding to users' requests, internet search engines are relying more and more on criteria that are not directly deducible from web resources themselves, but that allow to capture some information about the publishers and the users of the website.

In this article, examples are presented, which show the pragmatic criteria adopted by some internet search engines in the three main phases of their working: spidering, indexing and responding.

Keywords

Websites, internet, search engines, pragmatics, ranking algorithm, communication, spidering, responding, indexing.

A Communicative Approach to Web Communication: the Pragmatic Behavior of Internet Search Engines

1. INTRODUCTION

In the last years, the need for an approach to website communication that takes into account not only websites' technological aspects, but especially their communicative features clearly emerged (van der Geest 2001). Thus, to depict a comprehensive map of what a website is, we need a complex model which could account both for the various dimensions in a synchronic perspective and for the processes required to project, build-up, run, maintain, promote and evaluate a website (diachronic perspective).

The Website Communication Model – WCM (Bolchini et al. 2004; Cantoni & Piccini 2004; Cantoni & Tardini 2006) – provides such a map, in that it helps to distinguish five main areas of interest (see Figure 1):

1. those of *contents and services* offered through a website;
2. *accessibility tools*, i.e. the tools needed to access contents and services, with the related technological and graphical issues;
3. *publishers*, with the issues of website projecting, planning, running, promoting and maintaining;
4. *users*, with the issues of usability, web promotion and access analysis;
5. the *ecological context* of a website (i.e.: its relationships with the web as a whole).

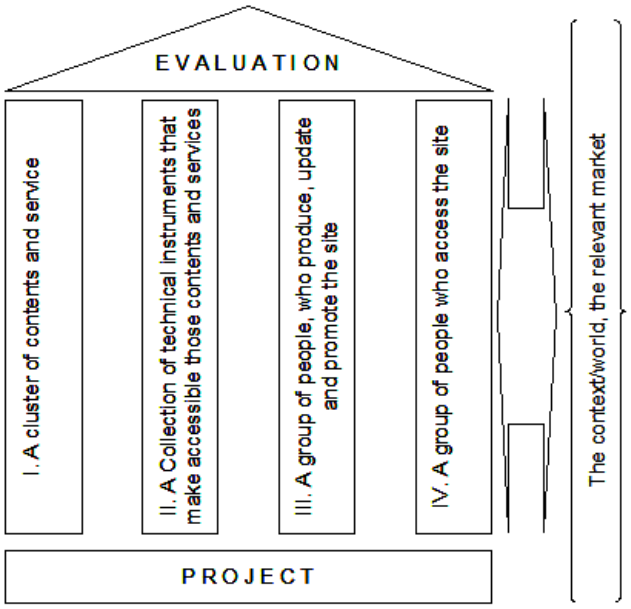

Figure 1. The Website Communication Model (WCM).

Within the WCM one of the main activities related to the publishers and users pillars is website promotion. Internet search engines are one of the most powerful tools for online promotion; in particular, they are very important in order to catch users' first visits, i.e. in order to get new clients access the website the first time (Middleberg 2001).

2. HOW INTERNET SEARCH ENGINES WORK

The huge amount of information available in the internet makes it difficult to communicate effectively both for websites' publishers and users: on the one hand, publishers need to make their websites visible and standing out in the mass of available information, on the other hand users need to easily find relevant information without getting useless ones.

Search engines are services that allow users to make full-text searches on the content of web pages; basically, they consist in big databases that archive web pages, index them and present them to users depending on their requests.

According to the method used to gather information, three different types of search engines can be singled out: 1) crawler-based engines, which are powered by spiders; 2) human-powered directories, where the submission of information relies on humans; 3) a combination of crawler-based and human-powered search engines.

Directories are big archives where websites are classified in a tree structure: every website that enters the directory is assigned to one (or more) category or sub-category. Ideally, categories should be exhaustive, i.e. they should cover all human knowledge, and should be reciprocally exclusive, i.e. one category should not overlap with another.

Directories have two main characteristics: 1) they are managed by human editors, who decide whether or not to insert the websites in the directory's database and – if yes – decide to which category (or categories) it is to be assigned; 2) they index websites, and not single web pages.

In spite of their success in the first ten years of the web, directories are not always the most suitable tool to categorize websites, first because it is often difficult to respect the rule of reciprocal exclusion; links among categories are put to try to prevent inefficient searches, but yield to confusion. Furthermore, a strict classification, as directories are, is an enforcement: in the offline world these limitations are necessary due to lack of physical spaces – shelves in a library are a typical example – but in the online world there is no shelf.

An attempt to overcome these limitations might be seen in the spread of so-called *folksonomies*, as explained in (Shirky 2005). A folksonomy is defined by users that assign one or more tag (a label) to describe the website they want to classify. When the application domain is the web, folksonomies are particularly effective, being the web a large corpus of unstable entities without formal categories. A huge number of users guarantee a great quality in the definition of the

folksonomy, even if they are not coordinated and are not expert cataloguers, what could mean a bad categorization for a single user. The websites del.icio.us (del.icio.us) and Flickr (www.flickr.com) are the living examples that folksonomies may really work: in the first case, thousands of users tag the web pages they visit; the second is a website for sharing photos that uses tags to categorize them.

Let us leave directories, going back to proper search engines.

The general working of internet search engines can be divided into three main activities: 1) spidering; 2) indexing; 3) responding (see Figure 2).

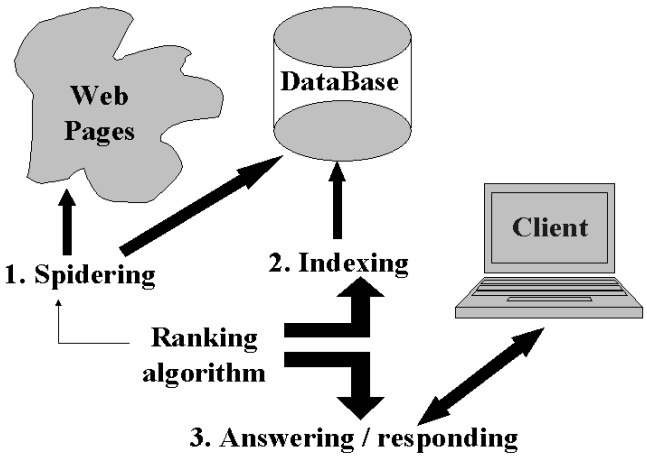

Figure 2. How internet search engines work.

2.1 Spidering

The first activity of an internet search engine is that of gathering web pages to create a database of web resources.

Spiders, or *web crawlers*, are robots (i.e. pieces of software) that surf the web in order to find web pages to be inserted into the search engine's database. Spiders go through the web according to given instructions, following links and fetching web pages to feed the database. Periodically,

the spider goes back to the same site in order to check any new information, changes and updates that it could present.

Some search engines allow websites' publishers to submit on their own initiative websites and web pages to their database. The submission of websites and web pages to a search engine can be made automatically or manually.

2.2 Indexing

Once the information pieces are loaded in the database, they have to be indexed in order to be made available for users' requests. Web resources are indexed on the base of a ranking algorithm, which controls the ranking of presentation of the resources to the users' requests. Every criterion used for ranking is aggregated into a unique indicator: the position on the results page.

The ranking algorithm used to index the web resources and present them to users' requests varies from one search engine to the other, but relies basically on two kinds of criteria:

1) criteria based on *intrinsic factors*: they are elements that are deducible from the web resources themselves, such as their URL, the name of the website, the titles of its pages and other information deducible from the source code with its tags and meta tags (meta tags are hidden HTML tags that provide information about the page, such as title, description, author, keywords);

2) criteria based on *extrinsic factors*: they are those elements that can not be found at all in the web page source code or in its URL, elements through which it is possible to capture some information about the publishers and the users of the website. Extrinsic factors are very useful, because they help taking into account the third, fourth and fifth dimensions of a website in the

WCM, i.e. the behaviors of people who manage and use the website, as well as the context where the website is in.

2.3 Responding

The third phase of the activity of an internet search engine is more concerned with the users' side: it is the phase of responding to the user's requests or searches. Also the activity of responding is based on the ranking algorithm of the search engine, since the visualization of the information given to the user's request depends on the ranking algorithm used.

What does it happen when a user types in the provided field box of a search engine the keywords s/he is interested in and gets back in a very few seconds a list of results? The search engine looks in its database for all the documents that match the keywords, finds all the related ones, and presents them to the user in an established order, according to its ranking algorithm. It is worth reminding here that the user's search does not actually take place over the internet, but s/he searches through the index created by the search engine, i.e. through its database.

3. THE PRAGMATIC TURN OF INTERNET SEARCH ENGINES

After the first 'ludic' years of the web, an important turn occurred in the functioning of internet search engines: borrowing the term from the linguistic tradition, we call it the 'pragmatic' turn of search engines.

3.1 Syntactics, semantics and pragmatics

In 1938, semiotician Charles W. Morris distinguished three branches within the semiotic field: syntactics (or syntax), semantics, and pragmatics. Being semiotic the science of signs, Morris defined syntactics as the study of "the formal relation of signs to one another", semantics as the

study of “the relations of signs to the objects to which the signs are applicable”, i.e. their *designata*, and pragmatics as the study of “the relation of signs to interpreters” (Morris 1971).

3.2 The need for a pragmatic turn of internet search engines

Internet search engines are taking into account more and more the behavior of people who publish and use websites, while at first they focused their attention almost exclusively on some syntactic and semantic features of web pages. The issues concerned here are the criteria followed by search engines to regulate websites submission to their indexes and to present the results of a query to their users.

At first, search engines usually allowed for free website submission, trying to compete on the field of completeness; their main objective was to have in their database as more web pages as possible, in order to be sure of offering to their users all the possible resources that matched their queries. However, the huge results they gave to almost every query was becoming more and more a problem for their users, configuring a situation that many call *information overload*: users got so many resources and so many documents back from the search engine that they were flooded with information and were not able to understand which ones were really relevant and useful to them and to select them among the others.

For this reason, search engines started to put many restrictions to the possibility for a website to be indexed. This change is due to the fact that if a high number of indexed pages helps fulfill the need for *recall* – all the pages that meet a given query are indexed – it reduces at the same time the chance for *precision* – only the relevant pages are presented to the user, and in a proper order (ranking).

Coming back to the abovementioned distinction, criteria based on intrinsic factors rely mainly on *syntactical* features of the indexed pages, i.e. they rely on a formal correspondence between signs, namely the keywords typed in by the users and some textual elements contained in the indexed pages; or on *semantics*, as long as meta tags provide a trustful semantic information about the actual resource's content. Criteria based on extrinsic factors rely on pragmatic elements, i.e. on elements that do not concern directly the content of the pages, but mainly the context where they are used, in particular the behavior of the publishers and of the readers of the web pages. These elements can provide some information about the real interest and the real motivation of publishers, by assessing, for instance, how often they update their pages, how much they are willing to pay to have actual communications on their websites, and so on; they can also provide some information about the readers' interest for a resource, by assessing, for instance, the popularity of a resource in the community of its users.

Thus, search engines are trying to take more into account extrinsic (pragmatic) elements, which help to assess not only a formal correspondence between queries and indexed web pages, but also the actual communities behind the web resources they have indexed: that of publishers and that of readers (users). In other words, it is clearly recognizable in the evolution of search engines a shift from purely syntactic to pragmatic criteria for the indexing and the presentation ranking of web pages. Both strategies, it is to be underlined, have the same goal: that of better matching – semantically – users' queries and search engines' answers.

4. PRAGMATIC STRATEGIES OF INTERNET SEARCH ENGINES: SOME EXAMPLES

If we go back to the search engine schema, we can find pragmatic strategies in all the three main activities done by an internet search engine. Let us present them in the same order.

4.1 Pragmatics in spidering

To improve the quality of indexed web pages, a search engine can decide to reduce the number of spidered items – according to certain criteria. In particular, the most adopted strategies are:

- not allowing automatic submissions;
- accepting (only) paid submissions.

Both are targeted at assessing the senders' *commitment*: are they really interested in having their web pages visited by the search engine users?

In the first case – stopping automatic submissions – the search engine does not ask for money, but for time: to feed a new resource one has to demonstrate s/he is a human being, who is devoting his/her time to this.

Money is a quite clear testimony of commitment, although a gross one. So some search engines ask for a payment in order to be spidered, or to be spidered on a given frequency in time: the idea behind it is that if you pay to be in a search engine, you must have something interesting to say.

Also directories have adopted this strategy: in this case, if one wants his/her website to be considered in a given period, has to pay. Again, this is a strategy to pre-check (indirectly) the quality of a website through the commitment of its publisher.

4.2 Pragmatics in indexing

While every ranking algorithm has to take into account computational linguistic rules (Zampolli 1998) – to match keywords, and to assess their relative relevance in a given corpus –, it can also embed pragmatic rules, to ensure a higher level of relevance.

The most used pragmatic strategy here is that adopted by Google: the so called *link popularity*. Hereafter how it is explained by Google itself: "PageRank [the ranking algorithm used by Google] relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important." Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Of course, important pages mean nothing to you if they don't match your query. So, Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query" (http://www.google.com/technology).

Actually, link popularity succeeded to reinterpret the apparently flat link structure of the world wide web as a hierarchical one, looking for an automatic strategy to reconstruct a hierarchy of sources (Gackenbach & Ellerman 1998). To do so, the web is considered as a system itself, and

not as being just a casual collection of pages. *Blogs*, as websites with high density of links, are a precious resource to evaluate link popularity that evolves in real time.

It has to be stressed that link popularity is a double indicator: it indicates explicitly a judgment of interest – usually a positive one (like a "vote") – done by the person who publishes a website toward another website, but it indicates also, inferentially, paths of actual usages: the many backlinks a website has, the more visits it is likely to receive. The ranking algorithm can also use the active part of an external backlink to assess the content of a given page. It is not likely that external links lie on the content of the target linked page. Moreover, in order to decide whether or not to select a link, one's judgment lays mainly onto the content of the link itself, hence, when adding a link, publishers are likely to word it in a suitable format, to enable sound choices by readers.

If link popularity infers usages, *click popularity* – another strategy implemented by search engines – measures them. Click popularity has been used to correct the result of the ranking algorithm through the feedback given (involuntary) by search engine users. Let us pretend that all the users entering the keywords "Cristoforo Colombo" do not click on the third result offered by a search engine: it can measure their clicks – and the time they spend on a given website before coming back to the results' page of the search engine – and use them to correct/integrate this feedback to better its ranking algorithm.

Another use of click popularity is done, for instance, by search engines which offer a *pay per ranking* service. If a given item is not clicked by users, it is discarded, even though it has a good bid. This approach matches quite well the interests of a search engine – if people do not click on

items, they do not get paid – and those of its users: if they do not click, it means that an item is not relevant for them; hence search engines ensure a better service to their users by removing it.

Another extrinsic element that can be embedded into the ranking algorithm is *time/currency*: resources which are more frequently/recently modified can be considered of higher quality than those published earlier.

Money is used also as a pragmatic indicator by ranking algorithms. A website's publisher can bid on given keywords, so that the webpage s/he submitted has a good position when those keywords are submitted to the search engine (e.g.: www.overture.com). In this case, the rank is just based on the amount of the bid: the more you bid, the higher you go; in case of the same amount, the search engine could award the better position to the webpage that was submitted before (which means, again, that it paid more…).

Integrating bids into a search engine algorithm has an intrinsic limit: only the best bidders can take advantage from it: if 100,000 people bid on the same keywords, only items which end up in being in the first page usually are selected, while for all the others bidding becomes simply useless. Due to this aspect, the struggle for getting the first positions is quite high, and items in top positions are frequently exported through syndication agreements.

4.3 Pragmatics in responding

When answering a user's question, a search engine can take into account contextual items, hence customizing results according to explicit and/or implicit indications given by the user.

In particular, implicit information can be inferred about the user's language and nation, so that the search engine offers a specific interface. There are some experimental attempts to integrate in the SERP (Search Engine Results Page) explicit information given by web surfers. For instance,

Outfoxed (getoutfoxed.com), a Firefox plugin, modifies the Google SERP on the basis of the rating of a given community by showing a mark about the relevance and the security of a website and re-ordering the results.

Search engines that index also news items or blog posts consider the moment in which they are published, selecting only the more recent leads.

Previous customizing choices done by the users can be taken into account by a search engine, hence "packing" results according to its users' requests. This is only limited to some GUI features (language, number of results visualized on one page). Some search engines (e.g. Google, Yahoo!) offer a more personalized page where users may see news headlines, weather, quote of the day and/or their email inbox. Actually, no major search engine offers real personalized search, intending with 'personalization' the fact that the ranking of results is calculated on user profiles.

While answers based on geographic information are usually explicitly elicited by users, through their entering a reference in space when doing a search, or through a customization choice, mobile technologies are opening huge possibilities to fully and transparently integrate user's spatial coordinates (e.g. through geo-localization) into the elements a search engine considers when compiling its answers. Google is also offering answers via SMS.

Some search engines divide vertically their answers into logical sections such as news, shopping, blogsearch, groups, travel. These may result from a precise choice by search engine managers or automatically, from clustering. *Clustering* is a technique to group results pages with similar contents. For example, searching for "lugano" on a popular cluster engine (such as Mooter, www.mooter.com), could yield to following clusters: lake, university, hotel, city, switzerland,

casino and others. Clustering can be seen as an attempt to take into account the context of a web page, in that it considers the web page as inserted in the whole "world" represented by all the results of a given search.

A search engine that is trying to verticalize results very strongly is A9.com. It does not lack in immediacy (type and go) but uses more tabs for a single search: web, movies, images, references (with lot of reference sites, dictionaries, wikipedia) and others.

In search engines, the SERP has not been changing for many years; thus, lot of pragmatic improvement could be done in the field of the visualization of the results. For example, Exalead (www.exalead.com) shows visual previews (as thumbnails) of the results found. There are also some browser plugins that enhance the Google SERP by adding thumbnails or popups with a preview of the linked page. See, for instance, Browster (www.browster.com) or LostGoggles (lostgoggles.com).

More complex improvements, such as changing the usual list, involve the abovementioned clustering of results. Kartoo (www.kartoo.com) uses clustering behind the scenes and provides the results' pages with a flash tool for surfing clusters.

Major search engines dare not to introduce big changes, but some of them offer little improvements. Google, for instance, verticalizes the results' pages with suggestions such as flights and news, calculator and money conversion.

5. CONCLUSIONS

The abovementioned examples clearly show the pragmatic aspects of web communication internet search engines are taking more and more into account. As a matter of fact, syntax is not sufficient in order for search engines to give enough relevant results to their users, nor is

semantics: on the one side the formal correspondence between the signs produced by the users (the keywords they search for) and those used by search engines to index their web resources cannot guarantee the quality of the results offered by search engines, due to a lack of relevance; on the other side, neither the exact correspondence between the keywords used to index the web resources and their real content can be guaranteed, thus causing once again the presentation of many irrelevant results to the users.

In order to cope with this problem, internet search engines are trying more and more to rely upon pragmatic features of websites, i.e. they are taking into account the behaviors of people who publish a website and people who visit it. This turn can be traced back to the growing awareness that websites – and, broadly speaking, electronic communication – are used by real communities of persons in order to fulfill real communicative needs. In other words, the pragmatic turn of internet search engines fits in the more general development of web communication, which passed from the reflection on the pure technical possibilities allowed by digital tools to the observation of the real uses of electronic texts, i.e. on the consideration of the publishers' and users' intentions.

6. REFERENCES

Bolchini, D., Arasa, D. & Cantoni, L. (2004), "Teaching Websites as Communication: A 'Coffee Shop Approach'", in L. Cantoni & C. McLoughlin (eds) *Proceedings of ED-MEDIA 2004*, Norfolk, VA, AACE, pp. 4119-4124.

Cantoni, L. & Piccini, C. (2004), *Il sito del vicino è sempre più verde*, Milano, FrancoAngeli.

Cantoni, L. & Tardini, S. (2006), *Internet (Routledge Introductions to Media and Communications)*, London – New York, Routledge.

Gackenbach, J. & Ellerman, E. (1998), “Introduction to Psychological Aspects of Internet Use”, in J. Gackenbach (ed.), *Psychology and the Internet: Intrapersonal, Interpersonal, and Transpersonal Implications*. San Diego, CA – London, Academic Press, pp. 1-26.

Geest, T. van der (2001), *Web Site Design is Communication Design*. Amsterdam – Philadelphia, PA, John Benjamins.

Middleberg, D. (2001), *Winning PR in the Wired World. Powerful Communications Strategies for the Noisy Digital Space*, New York et al., McGraw-Hill.

Morris, C.W. (1971), “Foundations of the Theory of Signs”, in C.W. Morris, *Writings on the General Theory of Signs*, The Hague, Mouton.

Shirky, C. (2005), *Ontology is Overrated. Categories, Links, and Tags*, http://www.shirky.com/writings/ontology_overrated.html (retrieved: 20.3.2006).

Zampolli, A. (ed.) (1998), *Survey of the State of the Art in Human Language Technology*, Cambridge, Cambridge University Press.