Navigating for Intelligence Islands through the Information Ocean

Memoria di licenza
di
Doojin, SONG
Matr. Nr. 99 - 980-146

Relatore: Prof. Mehdi, JAZAYERI

Anno Accademico: 2003-2004
**Acknowledgement**

"Teachers open the door, but you must enter for yourself."
- Chinese proverb

A special thanks to Prof. Jazayeri for his important advices, untiring patience, and encouragements.
Preface

“I have no special talents. I am only passionately curious”
- Albert Einstein

The reason to write this thesis is based on my curiosity about “how can the Internet search agents be an improvement for the quality of the human decision making intelligence? Since, the useful information in the Internet is extended like an ocean”.

The definition of Rogers, E.M. and Kincaid, D.L. (1981), “Communication is a process in which participants create and share information with one another in order to reach a mutual understanding” is the landmark and the starting point for the communication approach of the thesis.

To present a more complete picture of the topic, I will discuss also some others related communication and information theories. After that, the Information risks and the Internet search engine’s state of art will be examined. This will allow me to focus on the main problematic aspects of the use of search engines experienced by users. Consequently, these Internet search engine user’s major difficulties to use it will be explored within the survey.

Lugano, May 2004

Doojin SONG
Table of Contents

Part I - Introduction
1 Homo Communicatio
2 Computer Mediated Human Communication
3 Internet Search Engines & Communication

Part II – Communication & Information Theory
4 Communication Theory
5 Information theory

4.1 Communicatio
4.2 Shannon & Weaver’s Information Theory
4.3 Source-Message-Channel-Receiver model
4.4 The notion of interpretation
  4.4.1 Message & Meanings
  4.4.2 Empathy & Meanings
4.5 Credibility of Communication
4.6 Helical Model
4.7 Communication as a Convergence process

5.1 Information definitions
5.2 Uncertainty and Information
5.3 Cognitive load
5.4 Information Risk
  5.4.1 Information greed
  5.4.2 Information overload
  5.4.3 Misinformation on the Internet
### 5.5 Information Retrieval & Internet Search Engine

- 5.5.1 Information Retrieval ................................................................. 30
- 5.5.2 Internet Search Engine & Personalization .............................. 33

### Part III – Internet.................................................................36

#### 6 The state of art of the World Wide Web ..............................36

- 6.1 Beginning of the WWW ...............................................................36
- 6.2 Nature of World Wide Web ........................................................39
  - 6.2.1 How is the Web organized? ..................................................... 39
  - 6.2.2 Connectivity of the web .......................................................... 42
- 6.3 The quantity view of World Wide Web ..................................43
- 6.4 Categorization of web .................................................................44
  - 6.4.1 Deep Web ............................................................................. 45
  - 6.4.2 Fixed or Surface web .............................................................. 46
- 6.5 The user statistics ....................................................................46
- 6.6 Gateway to the Internet .............................................................48

#### 7 Internet Search Engines .........................................................49

- 7.1 The status of search engines ....................................................50
  - 7.1.1 Crawler - Based Search Engines .......................................... 50
  - 7.1.2 Directories ............................................................................ 51
  - 7.1.3 Meta-Search Engines .............................................................. 52
- 7.2 How many internet pages have been crawled? .......................54
- 7.3 Query matching and Results representation ..........................55
  - 7.3.1 Basic concept ........................................................................ 55
  - 7.3.2 Life of an Internet searching query (i.e. Google.com) .......... 56
- 7.4 Search results presentation technology ......................................57
  - 7.4.1 The relevance based methods ................................................. 57
  - 7.4.2 Clustering ............................................................................. 58
  - 7.4.3 Visualization .......................................................................... 60
8  Available data on Internet & Search Engines.................................63

8.1  How do Internet users behave when searching?.........................63
    8.1.1  Information flow pursuing pattern........................................ 63
    8.1.2  Query statistics................................................................. 64

8.2  Querying pattern study...............................................................65
8.3  Paid search results ......................................................................66

Part IV – Empirical Study...................................................................68

9  Survey on information overload & searching behavior....................68

9.1  Motivation.....................................................................................68
9.2  Relevant results ...........................................................................68
9.3  Some other results ......................................................................72
9.4  Limits of survey...........................................................................77

Part V – Work in progress...................................................................78

10 New Approaches towards High Precision......................................78

10.1  Semantic web ...............................................................................78
10.2  Personalization............................................................................80

User’s Searching Pattern Study.........................................................81

Part VI – Conclusion ...........................................................................84

BIBLIOGRAPHY..................................................................................86

ANNEXES ............................................................................................99
| Figure 4-1 | Shannon & Weaver Communication Theory | 14 |
| Figure 4-2 | Berlo's SMCR Communication Model | 15 |
| Figure 4-3 | Helical Model | 20 |
| Figure 4-4 | Basic components of the Convergence Model of Communication | 22 |
| Figure 5-1 | Hierarchy of Knowledge | 24 |
| Figure 5-2 | Recall and Precision graph | 32 |
| Figure 6-1 | Timeline of World Wide Web | 38 |
| Figure 6-2 | Connectivity Model | 42 |
| Figure 6-3 | Number of Hosts | 44 |
| Figure 6-4 | the Surface web & the Deep web | 45 |
| Figure 6-5 | Composition of Fixed web pages | 46 |
| Figure 7-1 | Number of crawled pages by search engines | 54 |
| Figure 7-2 | Time lines of search engine’s crawled pages | 54 |
| Figure 7-3 | From Query to answer’s life circle | 56 |
| Figure 7-4 | Google’s pagerank processes | 57 |
| Figure 7-5 | Vivisimo’s search results pages | 59 |
| Figure 7-6 | Kartoo’s search results page | 61 |
| Figure 7-7 | Groxis’ search results page | 61 |
| Figure 7-8 | TouchGraph’s interconnected link structure visualization | 62 |
| Figure 8-1 | Book use flow chart | 63 |
| Figure 8-2 | Web page use flow chart | 64 |
| Figure 8-3 | Searches Queries per day, February 25, 2003 | 64 |
| Figure 9-1 | Major difficulties to use Search engine | 69 |
| Figure 9-2 | Motives of getting lost in Internet | 70 |
| Figure 9-3 | Persistency to look for results | 71 |
| Figure 9-4 | Search engine use frequency | 72 |
| Figure 9-5 | Major usage of Internet | 72 |
| Figure 9-6 | Internet navigation access modes | 73 |
| Figure 9-7 | Major search engine used | 73 |
| Figure 9-8 | Frequency of query reformulation | 74 |
| Figure 9-9 | Query refining frequency | 74 |
| Figure 9-10 | Query refining used function | 75 |
| Figure 9-11 | Search engine use type | 75 |
| Figure 9-12 | Keyword saving at bookmark | 76 |
| Figure 9-13 | Useful link management modes | 76 |
1 Homo Communicatio

Rogers, E.M. and Kincaid, D.L. (1981) define “Communication is a process in which participants create and share information with one another in order to reach a mutual understanding”. In that sense, I dare to say that Communication is a vital element of society. When communication flows fluently, the society lives in a healthy state. Also the content is a vital element of the communication. The content could be translated as Information or data which is transformed to the Knowledge and the Wisdom (Cleveland, H.1982). Moreover, Information becomes precious when it has been communicated. Therefore, communication and information have a reciprocal relationship.

Nowadays, information acquires its assessment and the notoriety more than ever and it has become widely handy, a merit of Information Technology expansion. The easy access to information via Internet has been a cutting edge effect like the invention of print, which had permitted to create a great quantity of information.

In our communication epochs (McLuhan, M. 1964), in particularly, **Internet**¹ had been created for the purpose of transferring quickly or

---

¹ worldwide network of computer; computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange. Source: WordNet(2004) Internet is the example of internet. (Note: capital 'I'). The Internet is the largest Internet (with a small "i") in the world. Source: The Free On-line Dictionary of Computing, © 1993-2003 Denis Howe
instantly the digitalized files between closed communities from the diverse part of the planet. These communicational requisites of spread but closed communities have had an unexpected expansion impact into the whole world for disclosed and untied communities. Who has the possibility to connect to Internet has a possibility to getting to and to download the documents of others who published them in digital format. Hence, it is a time and cost effective advantage.

Beside of other advantages and disadvantages, the Internet is a symbol medium of the information age (McLuhan, M.1964). Internet coexists with other communication media; the ‘Homo Comunicatio’ is ‘Always up to date’.

‘Always up to date’ information might compromise the quality of the communication process through the increasing quantity of information to which digest, too. Consequently the communication quality could be decreased for the reason that mutual understanding becomes a hard task to reach.

2 Computer Mediated Human Communication

Using new communication media means creating a new social culture. It changes the lifestyle and the mentality of society. Using computer mediated human communication has changed the modern society visibly. Especially the diffusion of using Internet, as a new interactive medium of the human communication, has changed the whole globe’s culture. Because the interactive communication technologies facilitate multidirectional information exchanges (Rogers 1981). Rogers explains that “Internet is a typical of new communication technologies that are interactive in nature; a kind of interpersonal communication occurs via an electronic communication channel, rather that face-to-face”.

And more, the Interactivity is the degree to which participants in the communication process can exchange role in, and have control over, their mutual discourse (Williams, Rice, and Rogers, 1988) therefore the distinctive aspect of interactive communication technologies is “reciprocal interdependence” (Markus, 1990).

Also, the information into the Internet has a different figure from the classic figure of information presentation. It is digital instead of analogical. It allows changing their huge amount and varied content type, into the same and limited device. However it generated a great amount of information in the time. Therefore, the Internet environment becomes a sort of chaotic world, a harsh world. In spite of that the decentralized networks have a self-policing function. Even though no single authority is in charge of Internet, social norms are upheld (Rogers, 1995).

3 Internet Search Engines & Communication

Internet search engines are ports or gateways to the Internet. The Internet search engines are useful tools to find out the useful information that we are looking for, from the Internet which can be represented as the ‘Information Ocean’.

---

2 “Mutual discourse” is the degree to which a given communication act is based on a prior series of communication acts.

“Exchange of roles” means the empathic ability of individual A to take the position of individual B, and vice-versa.

“Having control” means the extent to which an individual can choose the timing, content, and sequence of a communication act, search out alternative choices, etc. (Williams, Rice, and Rogers 1988)

3 In the case of interactive innovations, not only do earlier adopters influence last adopters, but later adopters also influence earlier adopters, in a process of reciprocal interdependence (Markus 1990)
But the actual problem for the ‘Internet search engines’ user is that often we find out a lot of irrelevant information for our purpose. In this sense, Internet search engines might end up useless.

Different academic and non academic articles and tools have been published, to improve the actual Internet information search engines. But the Internet search engines’ problem does not only originate from a technical issue; in fact, it is also originated from the human information-related communication problem.

The opening questions for the thesis are:

“Which are the major communication barriers for the best quality of Internet Information Searching?
If the causes are detected, how Communication could be improved?”

In this research, I have the intention to explore the user’s demands for Internet search engines. Therefore, which Communication model might give contributions to improve, in communication sense, the quality of the Internet Information searching?

The first part of the thesis is about communication and information theories studies.
And the second part will focus on what different search engines services are offered to users.
The third part is focused on the user’s demands from the user’s internet searching behaviours statistics. Suggestions will conclude it.
4 Communication Theory

“Theories are not right or wrong, only appropriate or inappropriate given the circumstances and the nature of the phenomenon to which they are applied.”

- Figueroa, M. E. et al 2002

4.1 Communicatio

Peters, J. D. (1999) explained about the definition of the Latin word ‘commūnicātiō’, which means to impart, share, or make common. The key root of ‘commūnicātiō’ is mun, related to such words as "munificent," "community," and "meaning". In Latin, communicatio did not signify the general arts of human connection via symbols, nor did it suggest the hope for some kind of mutual recognition. Its sense was not in the least abstract: commūnicātiō generally involved tangibles.
In the course of times, communication has got its significance also in abstract. Communication is an implementation of social phenomenon change or a habit change. By means of communication, people construct a culture based society, create new social contact. Communication influences to people’s habit change, as Charles Sanders Peirce (1839-1913) mentioned, or manipulate people too. The impact of human communication is enormous on human civilization. In past times, communication has evolved and settled as an important asset of the modern society.

Communication is not something that one person does to another person. Communication is a continuous, ever-changing, circular, process of interaction. Over the years, many theorists have developed models of communication. Communication models give us a way to visualize, or make more concrete, an abstract process. They were modelling the complexities of communication. They help us to understand the important components of the communication process and the relationships among them.

I will now briefly look at the story of communication theories.
4.2 Shannon & Weaver’s Information Theory

Claude Shannon (1948)’s creation in the 1940’s of the subject of information theory has had an important and significant influence on mathematics, particularly on probability theory. But Shannon did his work primarily in the context of communication engineering.

Unfortunately, the semantic aspect of information plays no role in the theory. In the Shannon paradigm, information from a "source" (defined as a stochastic process) must be transmitted through a "channel" (defined by a transition probability law relating the channel output to the input). The system designer is allowed to place a device called an "encoder" between the source and channel which can introduce a fixed though finite (coding) delay. A "decoder" can be placed at the output of the channel. However, Shannon’s paradigm is a linear, one way process from sources to receivers. After that, when response was integrated in his model, it was considered like “knowledge of effects”.

Shannon’s work provides a crucial "knowledge base" for the discipline of communication engineering. Most significant is the fact that his theory showed how to design more efficient communication and storage systems by demonstrating the enormous gains achievable by coding, and by providing the intuition for the correct design of coding systems.
4.3 Source-Message-Channel-Receiver model

The **Source-Message-Channel-Receiver** model (SCMR model) advanced by David Berlo (1960), has served as another useful way to view the communication process.

![SMCR Model of Communication](Source: www.uky.edu/CommInfoStudies)

The model helps us to focus on the elements that must be operating effectively for communication to occur. The model helps us identifying the location of potential barriers to effective communication.

By combining this model with the Shannon & Weaver model, we can look at the elements of the communication process. However, Shannon’s and Berlo’s paradigms are linear, one way processes from sources to receivers. More comprehensive models also include as elements the following: **Feedback**, **Noise**, and **Communication context**. Subsequently, when feedback was integrated in these models, it was considered like “knowledge of effects” (Figueroa, M. E. et al 2002).
4.4 The notion of interpretation

Berlo (1960), focused on the ‘content of communication’. His paradigm is that “communication is the notion of interpretation”, one of the basic concept of semiotics and the meaning of communication.

4.4.1 Message & Meanings

Berlo (1960) mentioned that “Meanings are not in the message; they are in the message-users.” And more, Berlo explained “meanings do not reside within texts, but are actively constructed from those texts by readers”.

According to Berlo, we tend to interpret the world from our own vantage-point. This makes it difficult to communicate at all. It is often said that words do not mean the same to all people. It would be more accurate to say, as he proposes, that “words do not mean at all, only people mean”. Berlo says that “simply transmitting messages doesn’t ensure that we transmit meanings”.

Cherry, C (1980), explained that “even though people may say ‘the same thing’ (linguistically) on different occasions in conversation, each occasion, as an event, is observably different in many aspects from the other occasions; such differences depend upon people’s accents, their past experiences, their present states of mind, the environment, the future consequences of interpreting the message, knowledge of each other, and many other factors”.

Berlo deals with “the assumption implicit in transmission models of communication that meanings are somehow contained within the words we use”. And Berlo defines that “Communication does not consist of the transmission of meanings, but of the transmission of messages”.
Clearly, if the meanings of the signs we use were enclosed within signs, then we would be able to know any alien language without additional doubt. The elements and structure of a language do not themselves have meaning. They are only sets of symbols, signs that cause us to bring our own meanings into play, think about them, rearrange them etc.

Even if the elements of messages are received with 100% fidelity, there is certainly no guarantee that their meanings will be received. When we refer to the dictionary to find out the exact meaning of a word, often we find out different meanings for a word. **Meanings** were not somehow given when our codes were 'invented'; they arise out of a social process and therefore **change and evolve as society changes** and are different from one section of society to another.

### 4.4.2 Empathy & Meanings

We can have **similar meanings** only to the extent that we have had, or can anticipate having, similar experiences

Berlo (1960)'s point about 'anticipation' is important, though. We can certainly understand people's stories of experiences they have had and we haven't. To be effective communicators, we have to have **the skill of empathy**[^4]. Without this skill of empathy, it would be very difficult indeed to communicate effectively.

Meanings are never fixed; as experience changes, so meanings change. Many of our communication problems arise from the mistaken assumption that my message means the same to you as it does to me. Paradoxically, the dialogue is one of the primary media of surmounting reciprocal incomprehension.

[^4]: we have to be able to put ourselves in another person's shoes and see the world from their vantage point
4.5 Credibility of Communication

“Ethos is the credibility that the author establishes.”
-Aristotle

Why is the credibility necessary in human communication? The answer seems obvious. Without establishing the message sender’s credibility, its communication risks to fail with high degree of probability, soon or later. Consequently the credibility takes an important role of communication.

An obvious principle of human communication is that the transfer of meanings occurs most frequently between two individuals who are similar.

Lazarsfeld and Merton (1964) said that:

1. **Homophily** is the degree to which two or more individuals who interact are similar in certain attributes, such as beliefs, education, social status, and the like.
2. **Heterophily** is the degree to which two or more individuals who interact are different in certain attributes.

This physical and social nearness makes homophilous communication more likely. Such communication is also more likely to be effective. More effective communication occurs when two or more individuals are homophilous, or if not, when they have high empathy. Effective communication between two individuals leads to greater homophily in knowledge, beliefs, and explicit behaviour.
In an other occasion, Larson (1989), explained “If the message recipient feels that the communicator is a very dissimilar type of person, he or she will also assume that the message that person delivers is in disagreement with the recipient’s point of view. That message is perceived as less credible than if it had been delivered by a member of the recipient’s social group”. Also Hovland, Janis and Kelley (1953), mentioned that “An individual is likely to feel that persons with status, values, interests and needs similar to his own see things as he does and judge them from the same point of view. Because of this, their assertions about matters of which the individual is ignorant but where he feels the viewpoint makes a difference...will tend to carry special credibility”.

One of the most distinctive problems in the Internet communication is that the participants are usually quite heterophilous. Aristotle’s “On Rhetoric Theory of Civic Discourse” explains that credibility can be established by demonstrating three characteristics in writing: intelligence, virtue, and goodwill.

Therefore, **credibility** is an abstract and intangible aspect of human communication. It is highly connected with a receiver's will or attitude. A sender’s credibility is established by means of communication, which creates certain type of relationship. However the credibility itself is judged by the receiver.

### 4.6 Helical Model

Frank Dance's (1967) related communication to a helix. The bottom of the helix is small. As the helix moves upward it becomes larger. But movement up the helix is slow; a **process of circular, back-and-forth motion**.
At the beginning, communicators share but a small portion of themselves in their relationship. Through a slow process of giving and taking, their relationship develops as, little by little, they commit more of themselves. Dance’s helical model helps us to remember that the communication process is ever-changing, circular, expanding, and building upon the past.

![Figure 4-3 Helical Model](Source: Richard Nitcavic, Ball State University)

### 4.7 Communication as a Convergence process

Basic communication studies have been based on linear models of communication, the process by which messages are transferred from a source to a receiver. Such one-way view of human communication accurately describes certain types of communication; many kinds of communication do really consist of one individual. But other types of communication are more accurately described by a convergence model.

The communication receivers began to shift from the action of sources on receivers to the relationships among participants (Schramm, 1973), and to convergence within communication networks (Rogers & Kincaid, 1981). Nowadays, this relationship paradigm coexists with the on way action paradigm.
In Rogers and Kincaid’s convergence model (1981), "communication" is defined as “a process in which participants create and share information with one another in order to reach a mutual understanding”. Generally communication ceases when a sufficient level of mutual understanding has been reached for the task at hand. Several sequences of information sharing about a topic may increase mutual understanding but not complete it. Mutual understanding is never perfect.

Also the communication convergence model is a first process of the Communication Accommodation Theory (Giles et al 1991). They define convergence as “a strategy whereby individuals adapt to each other’s communicative behaviours”. And Convergence is a selective process and also based on attraction (Giles et al.1987). Intuitively, when communicators are attracted to others, they will converge in their conversation. Obviously also, the similarity facilitates the convergence.

The convergence paradigm has validity within the boundaries created by the dialogue itself. The convergence takes place since those who do not see an issue the same way nor agree with other participants tend to stop participating in the dialogue and the “drop out” of the group (Figueroa, M. E. et al 2002).

Rogers and Kincaid (1981) explained that “this definition implies that communication is a process of convergence as two or more individuals exchange information in order to move toward each other in the meanings that they give to certain events”. This convergence / network communication paradigm have primarily three key features: Therefore, I will limit to present this paradigm in detail, since it is outside the scope of my thesis.
The first is the communication that describes a process of dialogue, mutual understanding and agreement. The second is that it stresses the important role of the perception and interpretation on participants, which treats understanding in terms of a dialogue, ongoing cultural conversation. The third is the social change which is based on community dialogue and collective action. The social change is specified by the social outcomes as well as individual outcomes (Figueroa, M. E. et al 2002).

---

Figure 4-4  Basic components of the Convergence Model of Communication

---

5 Information and mutual understanding are the dominant components of the convergence model of communication. Collective action of any group is based upon information sharing, mutual understanding and mutual agreement.

Figueroa, M. E. et al (2002) stated that “When different points of view and beliefs arise (divergence), further communication is required to reduce the level of diversity (convergence) to the point where there is a sufficient level of mutual understanding and agreement to engage in collective action and solve mutual problems”.

The experience of “so that’s what you mean,” indicates a process in which participants gradually converge toward a better degree of mutual and shared beliefs (agreements). The dialogue constitutes a feedback process and leads to a “series of diminishing mistakes, a reducing series of under-and-over corrections converging on a goal” (Rogers & Kincaid, 1981). Convergence does not imply perfect agreement, only the direction of movement (Figueroa, M. E. et al 2002) and the relationships.
5 Information theory

5.1 Information definitions

Talking about information’s definition, Shannon C. (1948) defines “Information as something that reduces the uncertainty”. And Bateson G. (1979) defines “Information as still that which changes us”. Further more, Cleveland H. (1982) shows that the hierarchy of the wisdom starts with data at the bottom; information comes afterwards, then knowledge, and at the end, it could reach wisdom, even though if getting to wisdom is not a simple step from knowledge. Anyhow, the switch from raw data to information is an effort of communication activity. Generally, raw data is not structured. Information is a structured and context-related, valuable source in human life.

Figure 5-1 Hierarchy of Knowledge
(Source: "Information as a Resource", Harlan Cleveland)
A message carries information inasmuch as it conveys something not already known. The answer to a question carries information to the extent it reduces the questioner's uncertainty.

Roger and Kincaid (1981) state that “Information is a difference in matter-energy that affects uncertainty in a situation where a choice exists among a set of alternatives”. By differences in matter-energy we mean inked letters on paper, sound waves travelling through the air, or an electrical current in a copper wire. Information can thus take many forms, as matter or energy. A technological innovation embodies information and thus reduces uncertainty about cause-effect relationships in problem-solving.

5.2 Uncertainty and Information

The reality is that people everywhere have hard choices to make.

-Susan Podziba

Uncertainty is the degree to which a number of alternatives are perceived with respect to the occurrence of an event and the relative probabilities of these alternatives. Uncertainty motivates an individual to seek information. Uncertainty implies a lack of predictability, of structure. In fact, information is a means of reducing uncertainty.

Kuhlthau, C. (1993) explained that uncertainty is a cognitive state which commonly causes affective symptoms of anxiety and lack of confidence. “Uncertainty due to a lack of understanding, a gap in meaning, or a limited construct, initiates the process of information seeking”.
Thompson (1967) and Eveland (1986) defined “Technology as Information,” who suggested the technology’s and the information’s role on uncertainty reduction. Eveland (1986) suggests that “Technology is information and transfer is a communication process, and so technology transfer is the communication of information.” The idea is based on technology as a mean for helpful uncertainty reducing deed, in fulfil a determined goal. Technology as information is quite an unfamiliar definition; however, information could be presented as technology.

5.3 Cognitive load

The term "cognitive load" has been used in reference to Human Computer Interaction (HCI) issues. Cognitive load was used as a measure of the information processing effort that a user must pay out to take notice of the visual stimuli contained in an interface and comprehend its significance.

Cognitive Load Theory has been developed by educational psychologists and is documented by Sweller (1988). Learning structures (schemas) are used during problem solving. The psychologist Cooper (1996) explains that Cognitive Load Theory can be used to describe learning structures. Intrinsic cognitive load is linked to task difficulty, while extraneous cognitive load is linked to task presentation. If intrinsic cognitive load is high, and extraneous cognitive load is also high, then problem solving may fail to occur. When intrinsic load is low, then sufficient mental resources may remain to enable problem solving from any type of task presentation, even if a high level of extraneous cognitive load is imposed. E.g. an expert’s problem solving ability is higher than a beginner at any domain.
Also, the user friendly interface was assumed that users would prefer an interface design that requires a relatively low cognitive load and at the same time, can result in high user satisfaction. A self-reporting method was used to obtain individual users’ assessments of the cognitive load associated with a particular interface.

In Information Retrieval (IR), the concept of cognitive load rarely extends beyond the ideas presented by Miller, G.A. (1956). In Miller’s famous paper “The magical number seven plus or minus two”, a human’s capacity for processing information was explored. It was concluded that short-term memory (working memory) has a limited retention of seven plus or minus two similar objects.

Some of the more insightful studies in IR have shown that recognising the limitations of working memory may not be the only method of minimising cognitive load. Beaulieu (1997) suggested that there is a need to consider cognitive load not just in terms of the number and presentation of options, but more importantly to take account of the integration and interaction between them. Beaulieu (1997), however, was not the first, as Chang and Rice (1993) had proposed that interactivity could reduce cognitive load.
5.4 Information Risk

"Not everything that can be counted counts,
and
not everything that counts can be counted."
- Albert Einstein

5.4.1 Information greed

Are we forgetting more and caring less—or is there just too much to remember? More information has been produced and stored in the past 5 years than at any time in human history, in the form of e-mails, websites, cell phone messages, TV and print. Despite the fact that we are constantly bombarded with what seems like information, very little of it actually adds anything to human knowledge.

5.4.2 Information overload

Neuroscientist Keith Kendrick (2003) says, "We have got so much information now that you are likely to overload your own personal capacity. You have to be able to gloss over information extremely fast, instead of dealing with everything in phenomenal detail". He says that before the internet, the effort it took to get information, by going to the library, interviewing experts, etc., made the information more likely to stay in your mind because strong emotions became associated with it. "If you find it hard to get information, it sticks. Before the internet came along, just the achievement of finding out something was an emotional euphoria all by itself".
Talking about information’s definition, Teri K. Gamble & Michael Gamble (2002) describe “the situation that occurs when the amount of information provided by a speech maker is too great to be handled effectively by the listeners; the speech maker may provide more data than is necessary”. And Beekman G. (2001) defined that “The state of being bombarded with too much computer output; a hazard of the automated office”.

Wall Street Journal online published an article regarding the Information Overload of the managers, which said that “Just as feasting on an abundance of rich, available food can make you overweight, so we seem gorged on too much easily accessed information. We need a kind of information diet”.

People store up and remember either physical or non physical accordingly to their capacity. Why store or remember? The reason for doing it is that it would be useful for future use. Furthermore, the invention of Internet gears up the quantity of information enormously. In our small personal pc, there are intangible values of information stored too. Also, the increased speed of information transfer forced to invent digital Information management tools. Nowadays, the information quantity goes far beyond human natural information management capability to digest it. This is called Information overload. Schenk D. (1997) in argues that people don’t think deeply about the information they accept, particularly when exposed to a lot of information.

5.4.3 Misinformation on the Internet

The Internet is a powerful communication tool. Communicating over the Internet is relatively cheap; once connected, the vast majority of users can send e-mail, read from and post to newsgroups, and create
WebPages for free. The potential audience is huge and diverse, consisting of people from all over the world. Furthermore, information transfer on the Internet is fast; information can travel across the globe in a matter of minutes.

Unfortunately, misinformation travels as quickly and easily as everything else online. **Deciding what information and whose information to believe is hard**, because trivia information can be hard to distinguish. Internet users have learned to disbelieve at all or to distinguish some own criteria of judgement. However verifying information takes time and effort.

Compounding the problem is the amount of information available online. There is so much information available that the task of verifying everything a person sees is next to impossible. What’s worse is that there is no quality control. Thanks to the speed, scope and affordability of Internet communication, it is easy for misinformation to flourish and stick in the unconscious.

5.5 Information Retrieval & Internet Search Engine

5.5.1 Information Retrieval

A vast amount of information serves as a huge information repository. However, it also makes finding relevant information on the Internet extremely difficult. Helping users to find the information they require is the central task of most Information Retrieval (IR) systems or search engines.
While the problem of data and information retrieval has involved growing consideration, the reason for it is that there are great amounts of information to which accurate and speedy access is becoming continuously more difficult.

One effect of this is that relevant information might get ignored since it may be never uncovered, which in turn leads to greatly repetition of work and effort. With the arrival of computers, a great deal of thought has been given to how to use them to provide rapid and intelligent retrieval systems.

The three key components of an IR system from the definition of Rijsbergen Van (1979) are:

1. **Query representation**; representing the user’s information request
2. **Document representation**; representing the text collection
3. **Matching Function or ranking function**; ranking the documents according to their relevance

In a typical IR process, a user enters the IR system with an information request. The request is expressed as a **query** and submitted to the IR system. The IR system extracts keywords from the query, matches the query against the **document representations** in its collection, and assigns a **relevance status value** (RSV) to each document. The whole document collection is then ranked in descending order based on relevance status value and only the top $n$ documents are returned to the user.

---

6 Information Retrieval (IR)
The effectiveness of an IR system is normally measured by recall and precision.

1. **Recall** is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the whole collection.

2. **Precision** is defined as the number of the relevant documents retrieved to the total number of documents retrieved.

And when recall is low, precision is high.

![Recall and Precision graph](www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html)

For example, search engines’ results tend to get the low recall, consequently high precision. The reason of this strategy adoption is that while *irrelevant web site* that mentions user’s search terms may not be included, the results you seek are more likely to be among the first few returned. Since most searches don’t need to be exhaustive.

However, with the arrival of the Internet, new questions have been posed. One is whether techniques developed for small text collections can be extended to much larger collections. A recent study by Gordon and Pathak (1999) indicated that the precision and recall of commonly

---

7 [www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html](www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html)
used search engines such as AltaVista and Infoseek are usually very low. Users often have to filter through many web pages to find a small set of documents that satisfy their information needs.

The performance of an IR system can be affected by many factors (Chen, H., et al. 1998): the ambiguity of query terms, unfamiliarity with system features, factors relating to the document representation. Many approaches have been proposed to address these issues. For example, query expansion techniques based on a user's relevance feedback have been used to discover the user's real information need.

One more very important issue, which is a key to the performance of an IR system, is the ranking/matching function. The ranking function is used to order documents in terms of their predicted relevance to a particular query. However, people tend to have different criteria in terms of the interpretation of relevance for each retrieved document. A document relevant to person A may be totally irrelevant to person B for the same query, or vice versa. How to design a ranking function that is effective in most contexts is often related to the system designer's mental model or some other factors. In fact, it is very hard to design a ranking function that will be successful for every query, user, or document contexts.

On the other hand, I argue in favour of an approach that systematically adapts a ranking function and tailors it to the needs of different users. It could be a personalized search engine in a communication view. In particular, an inductive learning technique is used for the purpose.

5.5.2 Internet Search Engine & Personalization

An imminent problem faced by search engines is the achieving “High Precision, Low recall” by the personalization of their products and the
delivery of information related to the preferences of different users. Incorporating users’ preferences into the product design is of primary importance in the information age.

By personalization and learning the user’s profile on the relevance of documents, the same ranking function for a certain query can be used by the search engine again and again and save the knowledge worker time by eliminating the task of reviewing a large number of retrieval results.

Information Retrieval engineering treats questions democratically: all questions are processed by the same rules and operated in similar ways. Generally by their nature, questions carry more information. For this reason search engines often interpret and transform a user’s query during the retrieval process. These processes profoundly affect both the search results and the users’ ability to understand the relationship between their query and the returned results. Yet, few users are aware of those transformations or of the influence of those transformations on the sometimes mysterious results returned. Moreover, each search engine chooses its own combination of transformations, and might modify their choice without any indication of that change to the user. The application of these types of query transformations coupled with the common lack of feedback, leads users to develop inaccurate mental models of search engines. Mental models (Norman, D 1988) contain varying degrees of detail and accuracy; yet, they guide users’ operation, expectations, and understanding of tools.

Muramatsu & Pratt (2001) explained that when users’ mental models differ from the actual underlying system model, users often become frustrated and fail to accomplish their tasks. By transforming the user’s query without providing any feedback on those modifications, web search engines interfere with users’ formation of accurate mental
models and thus contribute to the users’ inability to find the desired information.

We search for information, and we anticipate that the information we want, will be available if we follow a certain pathway to it. Unfortunately it seems a complex goal to reach in short term, in spite of the effort of the search engines’ community research.
PART III – INTERNET

6 The state of art of the World Wide Web

6.1 Beginning of the WWW

The first recorded description of the social interactions that could be enabled through networking was a series of memos written by J.C.R. Licklider of Massachusetts Institute of Technology in August 1962 discussing his "Galactic Network" concept\(^8\). He envisioned a globally interconnected set of computers through which everyone could quickly access data and programs from any site. In spirit, the concept was very much like the Internet of today.

In the late 1966 Licklider went to DARPA (the U.S. Defence Advanced Research Projects Agency) to develop the computer network concept and quickly put together his plan for the "ARPANET", publishing it in 1967. At the conference where he presented the paper, there was also a paper on a packet network concept from the UK by Donald Davies and Roger Scantlebury of NPL (National Physical Laboratory, UK).

The objective was to develop communication protocols, which would allow networked computers to communicate transparently across multiple, linked packet networks. This was called the Internetting project and the system of networks which emerged from the research was known as the "Internet."

\(^8\) [http://web.mit.edu/invent/iow/cerf.html](http://web.mit.edu/invent/iow/cerf.html)
The system of protocols, which was developed over the course of this research effort, became known as the TCP/IP Protocol Suite, after the two initial protocols developed: Transmission Control Protocol (TCP) and Internet Protocol (IP). Scantlebury told Roberts about the NPL work as well as that of Paul Baran and others at RAND\(^9\). The word "packet" was adopted from the work at NPL and the proposed line speed to be used in the ARPANET design was upgraded from 2.4 kbps to 50 kbps.\(^{10}\)

In 1986, the U.S. National Science Foundation (NSF) initiated the development of the NSFNET which, today, provides a major backbone communication service for the Internet. With its 45 megabit per second facilities, the NSFNET carries on the order of 12 billion packets per month between the networks it links. The National Aeronautics and Space Administration (NASA) and the U.S. Department of Energy contributed additional backbone facilities in the form of the NSINET and ESNET respectively.

In Europe, major international backbones such as NORDUNET and others provide connectivity to over one hundred thousand computers on a large number of networks. Commercial network providers in the U.S. and Europe are beginning to offer Internet backbone and access support on a competitive basis to any interested parties.

"Regional" support for the Internet is provided by various consortium networks and "local" support is provided through each of the research and educational institutions. During the course of its evolution, particularly after 1989, the Internet system began to integrate support for other protocol suites into its basic networking fabric.

\(^9\) RAND: the name of which was derived from a contraction of the term research and development, focused on issues of national security. www.rand.org  
\(^{10}\) http://www.isoc.org/internet/history/brief.shtml
The present emphasis in the system is on multiprotocol interworking, and in particular, with the integration of the Open Systems Interconnection (OSI) protocols into the architecture.

Both public domain and commercial implementations of the roughly 100 protocols of TCP/IP protocol suite became available in the 1980’s. During the early 1990’s, OSI protocol implementations also became available and, by the end of 1991, the Internet has grown to include some 5,000 networks in over three dozen countries, serving over 700,000 host computers used by over 4,000,000 people.

The Coordinating Committee for Intercontinental Networks (CCIRN), which was organized by the U.S. Federal Networking Council (FNC) and the European Reseaux Associees pour la Recherche Europeenne (RARE), play an important role in the coordination of plans for government-sponsored research networking. CCIRN efforts have been a stimulus for the support of international cooperation in the Internet environment.¹¹

---

¹¹ source: [www.isoc.org/internet/history/cerf.shtml](http://www.isoc.org/internet/history/cerf.shtml)
6.2 Nature of World Wide Web

The analysis of the Web’s structure is leading to improved methods for accessing and understanding the available information, for example, through the design of better search engines, automatically compiled directories, focused search services, and content filtering tools.

6.2.1 How is the Web organized?

Web pages can be defined as hubs and authorities. A **hub** is a page that **points to many authorities**, whereas an **authority** is a page that is **pointed to by many hubs**. Characteristic patterns of hubs and authorities can be used to **identify communities of pages on the same topic**.

Unlike other great networks of the past century the Web is a virtual network of content and hyperlinks, with over a billion interlinked “pages” created by the uncoordinated actions of tens of millions of individuals. Because of the decentralized nature of its growth, the Web has been widely believed to lack structure and organization as a whole.

Pennock, D.M. et al (2000) indicates that the Web contains a large, strongly connected core in which every page can reach every other by a path of hyperlinks. This core contains most of the prominent sites on the Web. The remaining pages can be characterized by their relation to the core. In fairly large snapshots of the Web, these four components—core, upstream, downstream, and tendril regions—have roughly comparable sizes.
Moreover, the core is very compact: The shortest path from one page in the core to another involves 16 to 20 links on average, a “small world” situation in which typical distances are very small relative to the overall size of the system.

If the processes that drive Web growth are highly decentralized, then the power law must arise from a composite of local behaviour. This randomized, “rich-get-richer” process, the network grows by the sequential arrival of new nodes, and the probability that an existing node gains a link is proportional to the number of links it currently has. The result is a power law distribution of links. Deviations from power-law scaling occur, especially at small numbers of links. Furthermore, the deviation varies for different categories of pages. For example, the distribution of links to university home pages diverges strongly from a power law, following a far more uniform distribution.

Models of Pennock, D.M. et al (2000) seek to improve the accuracy of the original preferential attachment models. At a local level—the scale of small neighbourhoods and focused regions of the Web—the structure turns out to be even more intricate and quite non-uniform. Pages and links are created by users with particular interests, and pages on the same topic tend to cluster into natural “community” structures that exhibit an increased density of links.

Turning this observation around leads to a powerful method for analyzing the content of the Web. An unusually high density of links among a small set of pages is an indication that they may be topically related. A characteristic pattern in such communities consists of a collection of “hub” pages—guides and resource lists—linking in a correlated fashion to a collection of “authorities” on a common topic. A related pattern is one in which authorities on a topic link directly to other authorities, again creating a density of links.
Link analysis as a means of finding authoritative (Pennock, D. M. et al 2000), relevant sources on the Web has proven useful in the design of improved search engines. This application of link analysis has clear connections with, as well as interesting contrasts to, citation analysis of scientific literature and the identification of “central” individuals in a social network.

Knowing the characteristic link structures that identify Web communities, (Kumar, R. et al, 1999) one can examine a large snapshot of the Web for all occurrences of the link-based “signature” of a community. A community (Flake, G. W. et al 2000) can also be defined as a collection of pages in which each member page has more links to pages within the community than to pages outside the community. This definition may be naturally extended to identify communities with varying levels of cohesiveness.

As with the previous approach, this method of searching for communities reveals a remarkable degree of self-organization in the Web’s link structure, and textual analysis of the communities shows that the constituent pages are topically related. New or niche sites with few links to them may have difficulty competing with highly prominent sites for attention.

By favouring more highly linked sites, search tools may increase this effect. (Kleinberg, J. & Lawrence, S. 2001)
6.2.2 Connectivity of the web

This connected web breaks into four pieces.

The first piece is a **central core**, whose all pages can reach one another along directed hyperlinks - this "giant strongly connected component" (SCC) is at the heart of the web.

The second and third pieces are called **IN** and **OUT**. **IN** consists of pages that can reach the **SCC**, but cannot be reached from it - possibly new sites that people have not yet discovered and linked to.
OUT consists of pages that are accessible from the SCC, but do not link back to it, such as corporate websites that contain only internal links.

Finally, the TENDRILS contain pages that cannot reach the SCC, and cannot be reached from the SCC. Perhaps the most surprising fact is that the size of the SCC is relatively small.

One can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passaging through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE -- a passage from a portion of IN to a portion of OUT without touching the SCC.12

6.3 The quantity view of World Wide Web

Most magnitude studies of the World Wide Web have focused on the number of hosts connected to the network. There are 172 million (exactly; 171,638,297 in January of 2003) registered hosts. (A Host is, a computer system with registered IP address, advertised in the Domain Name Service)

The host numbers has experienced a growth of 130 % in quantity (which was in January of 1993; 1,313,000 hosts) in ten years.

Also there are 42 million (exactly; 42,298,371 in July of 2003) of registered web servers.

6.4 Categorization of web

From the statistics published in 2003 on the quantity of information in Internet, by the School of Information Management and System at Berkeley, it is possible to say the following.

The World Wide Web contains about 170 terabytes of information on its surface (fixed web pages) (e.g. 1 Terabyte: 50,000 trees made into paper and printed, or 10 Terabytes: The print collections of the U.S. Library of Congress); in deep web pages it contains 91,850 terabytes, in Email 440,606 terabytes.
6.4.1 Deep Web

Deep Web is the largest growing category of new information on the Internet. Bergman M.K. (2001) explained that “Deep Web sources store their content in searchable databases that only produce results dynamically in response to a direct request. But a direct query is a "one at a time" laborious way to search”. Therefore, “Without the directed query, the database does not publish the result” (Steve Burks, 2001).

Deep Web sites tend to be narrower with deeper content than conventional surface sites. The total quality content of the Deep Web is at least 1,000 to 2,000 times greater than that of the Surface Web. Deep Web content is highly relevant to every information need, market and domain. More than half of the Deep Web content resides in topic specific databases. A full 95% of the Deep Web is publicly accessible information not subject to fees or subscriptions.

A recent study\textsuperscript{13} revealed that the largest search engines today individually index at most 16% of the Surface Web. By missing the content on the Deep Web, Web searchers are searching only 0.03% - or one in 3,000 - of the content available to them.

![Figure 6-4 the Surface web & the Deep web](www.themesh.com/tr64.html)

\textsuperscript{13} www.themesh.com/tr64.html
6.4.2 Fixed or Surface web

From the School of Information Management and System at Berkeley’s research, the file types are grouped in 15 categories. The major file type is the image file and html and php files. Also the pdf files are quantified almost half of html files.

![Composition of Surface Web](image)

> Figure 6-5  Composition of Fixed web pages  
(Source:How much Information? 2003)

6.5 The user statistics

From the cyberatlas.internet.com statistics, in 2004 there will be about 945 million of online population, instead of 655 million (source: ITU) in 2002. The demographic distribution is as following roughly 30 percent are in North America, 31 percent are in Europe, and 31 percent are in Asia Pacific and 8 percent in other parts of the world. Also, the average
global Internet user spends 11 hours and 24 minutes online per month, according to Nielsen/NetRatings.

The Pew Internet and American Life Project report that on an average day, about 72 million United States users go online. And, most Internet users have searched for answers to specific questions. 83% of users have done a search online to answer a specific question, according to our latest survey on the subject in September 2002. The number of those who have used the Internet to answer questions grew of 24% from 79 million to 98 million between the fall of 2000 and September 2002. There are more searchers with higher education levels (college degrees) than among those who completed high school. Those with the most experience online and those with broadband connections are more likely than others to do online searches for information on any given day.

The act of using the Internet to answer basic questions is second only to email in its popularity as an online activity and easily one of the most essential functions of the Internet. Often one of the first activities that new users try, searching is an important step in how newcomers to the Web come to value the Internet as an everyday reference for finding useful information. Also, it is an important exercise in learning how to sift through the sometimes unreliable or irrelevant information that one can encounter during Web browsing. In the past, we have identified this activity as part of the early “search and learn” mode that allows new users to become comfortable with the fundamental workings and navigation of the Web.
6.6 Gateway to the Internet

The Internet is a system of storage and retrieval of information characterized by its enormous size, hypermedia structure, and distributed architecture. But entering into the Internet, it is necessary to know a gateway. It could be a search engines. When searching in the Web, a user can be overwhelmed by thousands of results retrieved by a search engine, few of which are valuable. Traditionally, in paper documents researchers have had clear-cut boundaries to look up information in an index, footnote, etc. On the web, we are sometimes limited by the novel characteristics of the tool for our search. We must first ask the question of how researchers look at a web search and do they browse for information.

The problem for search engines is not only to find relevant results, but results consistent with the user's information need (Barry, C.L. 1993). It is a user’s information need that determines which documents are valuable.

What happens if the user doesn't know what keywords to use? This situation can present itself when the user is searching new domains; when the user is uncertain about what to search for or when they do not know what keywords to use to get the desired results. There is a divide between the user’s search intent – their mind view of what kind of results they wish to find - and the effectiveness of the query they craft to match their intent.
7 Internet Search Engines

It can be roughly imagined that an information searcher into the Internet may be compared to a mushroom searcher into a forest. An expert on mushrooms knows well, first of all, about mushrooms, and about forests, climate, season, temperature. After that, they have knowledge about where they could find out it and a capacity to distinguish which of them they can collect or not. They consider all variables to find good mushrooms.

But the information searcher in Internet differs from the mushroom expert because of the necessity of a searching agent. That agent is called the Internet search engine. As a result, the Internet information searcher has to interact mostly with the Internet search agent. Consequently, knowledge of Internet searching agent may be required. This despite the fact that beginner searchers know that there is a variety of functions in search engines. Into a small dialogue box, the searcher submits a query and the search engines match that query within their own databases.

But finding out an exact answer from the search engines’ results, concerning on what the user request was, is often a tough task. Since the Internet area to search is vast, and unknown, the submitted keyword generates other context-related results, and often the searcher doesn’t know how to call the “mushroom” or has difficulty parsing in “...what I want to know about ...” in short terms or the search agent gives to the searcher too many similar answers. The first step (supposed that user at least knows what they are looking for) to interact with the search agent could be to know the Internet search engines’ nature.
7.1 The status of search engines

In the context of the Internet, the word "search engines" is most often used for search forms that search through databases of HTML documents gathered by a robot. The term "search engines" is often used to describe both crawler-based search engines and human-powered directories. The Internet is always growing and since these search engines search in different ways and search different parts of the Internet, doing the same search using different search engines will often give you differing results.

7.1.1 Crawler-Based Search Engines

Crawler-based search engines use automated programs, called robots, to search the web. A crawler is a program that downloads and stores Web pages, often for a Web search engine (Cho, J. Garcia-Molina, H. 2002). The crawler visits a web page, reads it, and then follows the links to other pages within the site. The crawler returns to the site on a regular basis, such as every month or two, to look for changes.

E.g. the AltaVista crawler is based on a large set of starting points accumulated over time from various sources, including voluntary submissions. The crawl proceeds in roughly a Breadth First Search\textsuperscript{14} manner, but is subject to various rules designed to avoid overloading web servers, avoid robot traps (artificial infinite paths), avoid and/or detect spam (page flooding), deal with connection time outs, etc.

\textsuperscript{14} A graph search algorithm which tries all one-step extensions of current paths before trying larger extensions. This requires all current paths to be kept in memory simultaneously or at least their end points.
Each instance of the AltaVista index is based on the crawl data after further filtering and processing designed to remove duplicates and near duplicates, eliminate spam pages, etc. (Broder, A. et al 2000)

After that, everything the crawler finds goes into the index. The index is like a giant file folder containing a copy of every web page that the crawler finds. If a web page changes, then this index is updated with new information. Sometimes it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a web page may have been "crawled" but not yet "indexed".

And the search engine software is the program that sort through the millions of pages recorded in the index to find matches to a search and rank them in order of what it supposes is most relevant.

7.1.2 Directories

Directories on the Internet were created following the model of traditional Directories since they are organized by category or topic, etc. About.com, Look Smart, and Yahoo are some examples of directories. Internet directories also now make use of search engine functionalities. Vice versa, many of the search engines (e.g. Google etc) offer directories of topics for those who prefer to browse.

The classification process could be roughly imagined as the postman who classifies the letters or other postal objects. When we write correctly the address of the receiver, the postal office worker could automatically classify the letters into a defined address classification.

Directories have the advantage over search engines of being very precise in the way they categorize pages. The advantages of directories are;
1. Collections of links to Web sites compiled by people, not software robots, sometimes experts in a subject.
2. Often carefully evaluated and kept up to date, but not always -- frequently not if large and general.
3. Usually organized into hierarchical subject categories.
4. Often annotated with descriptions.
5. Can browse subject categories or search using broad, general terms.
6. Searches need to be less specific than in search engines, because you are not matching on the words in the pages you eventually want.

Generally, in Directories, you are searching only the subject categories and descriptions you see in its pages.

What distinguishes directories from crawlers is that a web page subscriber submits a short description to the directory for their entire site, or editors write one for sites they review. A search looks for matches only in the descriptions submitted. Changing your web pages has no effect on your listing. Strategies that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The possible advantage is that a valuable site might be more likely to get reviewed for complimentary.

7.1.3 Meta-Search Engines

A meta-search engine is a multiple submitter to other search engines. It submits queries to other single search engines at the same time and gets back the results and visualizes the results to the user.
The idea behind a meta-search engine is that an Internet user should be able to visit one site, conduct a search, and view the results from multiple search engines on a single results page. The key difference from a general search engine is that meta-search engine does not create or maintain their own index of Web sites, they simply pass searcher's query on to other search engines and aggregate the results together for user’s convenience.

Many meta-search engines have special agreements with the search engines that they query for search results. These agreements allow them to access the search engine indexes through a "backdoor" and provide results at a more rapid pace than a human being that would conducting the searches on their own.

Additionally, a meta-search engine will sometimes process the results according to their own internal programs. For instance, they might eliminate multiple listings of a single site, apply the meta-search engine's own algorithm to the listings, "cluster" results together in easy to browse categories (e.g. Vivisimo), and create a visual map of the searched topic on the Internet (e.g. Kart00).
7.2 How many internet pages have been crawled?

At January 21st 2004, two well known search engines classified these numbers of webpage’s;

- Alltheweb: 3,151,743,117 pages crawled
- Google: 3,307,998,701 pages crawled

Figure 7-1  Number of crawled pages by search engines
(Source: searchenginewatch.com/reports/article.php/2156481)

And from the timeline of the webpage’s crawled we can notice each search engines’ efforts to have bigger own databases.

December 1995-September 2003

Figure 7-2  Time lines of search engine’s crawled pages
(Source: searchenginewatch.com/reports/article.php/2156481)
7.3 Query matching and Results representation

7.3.1 Basic concept

After that, robots have crawled and indexed, the search engines receive queries from users. Search engines do query **matching from own web pages data** and show up the search results. To show them, it is necessary to have an order. The criteria of order are defined from the search engine company technology. The majority of search engines show up most **relevant matched ones first**. Different search engines have different criteria of relevancy for standing (ranking) algorithms. Each ranking algorithm assigns different weights to different occurrences of the keywords, depending on where and in what form these matches are found.

E.g. one search engine uses rules relating matched location with keyword in the page and frequency method. Pages with keywords appearing in the title are assumed to be more relevant than others to the topic. The search engines also check if keywords appear near the top of web page, such as headline or first few paragraphs. It assumes that any page relevant to the topic will mention those words right from beginning.

The general sequence of matching scan order is;

1. Title
2. Keyword Density
3. Emphasized Text
4. URL
5. Popularity
Each algorithm generates the score needed to determine the ranking of the specific site. The higher the score of own relevancy criteria obtained from the applied algorithms, the bigger the chance for that specific site to appear in the search results.

7.3.2 Life of an Internet searching query (i.e. Google.com)

![Diagram of the life cycle of an Internet searching query](source)

The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result. The search results are returned to the user in a fraction of a second. The list, with abstracts, is displayed by the web server to the user, sorted (using a PageRank technology).
7.4 Search results presentation technology

7.4.1 The relevance based methods

7.4.1.1 PageRank™ by Google.com

![Figure 7-4 Google’s pagerank processes](http://www.pcmag.com)

PageRank relies on the nature of the web by using its vast link structure as an indicator of an entity 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 absolute 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. This is not an objective matching criteria of importance. Consequently important pages mean nothing to the user if they don’t match the user’s query. Importance is a quite subjective value and nuance value.
So, PageRank is combined with text-matching techniques to find pages that are both important and relevant to the user’s search. Hypertext-Matching Analysis: It analyzes all the content on each web page and factors in fonts, subdivisions, and the precise positions of all terms on the page. Google considers also the number of times a term appears on a page and examines major aspects of the page’s content (and the content of the pages linking to it) to determine if it’s a good match for the user’s query.

7.4.2 Clustering

“Science is organized knowledge.
Wisdom is organized life”
- Immanuel Kant

The need to organize large amounts of knowledge and information led to the development of classification schemes and other organizational tools. Classification schemes (Arlene, G. T. 1999) bring systematic order and control to the collection so that an information package can be retrieved according to a particular aspect of its character. Library classification schemes also improve retrieval through enhanced subject access via Online Public Access Catalogues.

A successful classification scheme is one that saves the time of the user by creating an order convenient to the user (Mortimer, M. 2000). Unfortunately, the name of information on the Internet is different from the library. Since search engines provide access and retrieval points for Internet resources, they have their own indexing organizational scheme. Therefore, search engines had been compared as a library for the Internet. However Internet search engines are not like a library.
The key difference between a library and an Internet search engine is the search engine's self constructed indexing or classifying technology. E.g. Google has its own method, which is different from Alltheweb's. Why did not search engines adopt the commonly diffused library classification? The advantage of the library classification would help much the consistency across varied information stored up by establishing a shared conceptual context (Albrechtsen, Hanne & Jacob, Elin. 1998).

One good example of clustering by a search engine company could be Vivisimo. Vivisimo is a meta-search engine, which offers an own clustering technique. As Miller (1956) said, this clustering technology could be good example to empower the span of attention.15

Figure 7-5  Vivisimo's search results pages
(Source: http://www.vivisimo.com)

15 “short term memory” can typically only hold 7 items +/- 2 - which they suggest affects our ability to analyse things and is limited by this short term memory, which psychologists now refer to as “working set memory”.
Clustering will help search engine users to empower “the span of attention”. Clustering makes easier the user’s assessment activity. The idea is to make several blocks depending on the similarity of each elements or scattered\textsuperscript{16} plain information. By clustering or grouping for their similarity on concern, user can discriminate easily than information which is treated at a same category for a keyword.

7.4.3 Visualization

\textit{A picture is worth a thousand numbers.}

-Unknown

Visualization of search results can be readied in different ways. However, in a dictionary the definition of visualization is “To form a mental image of; \textit{e.g. tried to visualize the scene as it was described}”\textsuperscript{17} With visualization, the user feels more comfortable than with a plain representation of idea. The possible explanation for visualization is that the user views simple text lists, and gets bored and overlooks possible information for a limits of the attention.

There is a good example of search results visualization. A meta search engine, Kartoo, or Groxis.

\textsuperscript{16} a haphazard distribution in all directions; move away from each other
source: \url{www.cogsci.princeton.edu/cgi-bin/webwn?stage=1&word=scatter}

Figure 7-6  Kartoo’s search results page
(Source from: http://www.kartoo.com)

Figure 7-7  Groxis’ search results page
(source: http://www.groxis.com)
This graph shows the web site’s link relationships.

Figure 7-8  TouchGraph’s interconnected link structure visualization
(Source: http://www.touchgraph.com/TGGoogleBrowser.html)
8 Available data on Internet & Search Engines

8.1 How do Internet users behave when searching?
8.1.1 Information flow pursuing pattern

It is simple to verify the reading pattern flow for a book. It is a sequential model, which means that the reader pursues a path of pages in order, following the editor's flow scheme. Excepting, there are many diversified personal reading patterns. This pattern can be found easy everywhere in our environment; films, theatres, TV dramas etc.

Figure 8-1 Book use flow chart
(Source: www.gl.iit.edu/llc/webworkshop/outline.htm)

However, the introduction of hyperlinks in Internet has consequences in handling paths. The interconnected mesh structure could be quite dispersive to the user and cause an unpredictable content flow scheme. In some cases, how the user can pursue depends on their need.
8.1.2 Query statistics

How many searches are performed each day? The following table shows self-reported figures:

<table>
<thead>
<tr>
<th>Service</th>
<th>Searches Per Day</th>
<th>As Of/Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>250 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>Overture</td>
<td>167 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>Inktomi</td>
<td>80 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>LookSmart</td>
<td>45 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>FindWhat</td>
<td>33 million</td>
<td>January 2003</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>20 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>AltaVista</td>
<td>18 million</td>
<td>February 2003</td>
</tr>
<tr>
<td>FAST</td>
<td>12 million</td>
<td>February 2003</td>
</tr>
</tbody>
</table>

There are roughly 625 million searches per day, by using popularly used search engines. It means that the online population does at least
one searching activity everyday. And according to the Nielson Netratings, statistics show searches per day as estimated by "search hours." It was then assumed that a typical search takes a user 20 seconds to perform.\textsuperscript{18}

From the statistics of America’s Online Pursuits, search engines have become an indispensable utility for Internet users. More than eight out of ten Internet users have searched the Internet to answer specific questions. Topics can range from the ridiculous ("How many times does my name come up on Google?") to the sublime ("Where was Buddha born?") to the heartbreaking ("My mom has breast cancer – I need information fast").

But the strategies are similar for all these questions – type keywords into a search engine and go from there. Internet searching is an umbrella activity that encompasses many of the other more specific activities.

8.2 Querying pattern study

The Keyword search approach is the same approach as with a word finding in a dictionary. The keyword search is a simple search method, a good search approach in spite of their limited nature.

In the usability study of 20 big websites by Jakob Nielsen, users had a success rate of only 64% in searching for - and finding - what they wanted.

He found also that users are somewhat unfortunate at query reformulation, and almost half the users whose first search failed gave up immediately. If they don't get good results on the first try, later search attempts rarely succeed. In fact, they often give up.

\textsuperscript{18} Source: adapted [http://www.searchenginewatch.com/reports/article.php/2156451](http://www.searchenginewatch.com/reports/article.php/2156451)
Their search success rate was: first query 51%, second query 32%, and third query 18%. In other words, if users don’t find the result with their first query, they are progressively less and less likely to succeed with additional searches. Many users don’t even bother.

The possible alternative could be a user re-education programs for the query art in search engine. But this is a quite vast goal and somehow unrealistic.

Realistically speaking, search engines designers should assume that most users won’t be willing or able to refine their queries.

Another reason to emphasize early success is that users typically make very quick judgments about a website’s value based on the quality of one or two sets of search results. If the list looks like junk, they may abandon the site completely. At a minimum, they will miss the site’s search in favour of external search engines like Google.

8.3 Paid search results

Another user study by Leslie Marable find out the following:
Most participants had little understanding of how search engines retrieve information from the Web or how they rank or prioritize links on a results page. The majority of participants never clicked beyond the first page of search results as they had blind trust in search engines to present only the best or most accurate unbiased results on the first page.

As a result, two-in-five links (or 41%) selected by our participants during the assigned search sessions were paid results.

Once enlightened about pay-for-placement, each participant expressed surprise about this search engine marketing practice. Some had negative, emotional reactions.

19 www.consumerwebwatch.org/project/staffbios/marablebio.html
All participants said paid search links on search and navigation sites were often too difficult to recognize or find on many sites, and the disclosure information available was clearly written for the advertiser, not the consumer. Search engine sites that were perceived to be less transparent about these related disclosures lost credibility amongst this group of online consumers.
PART IV – EMPIRICAL STUDY

9 Survey on information overload & searching behavior

9.1 Motivation

I wished principally to verify the Internet search engine user’s main claim of experiencing is the “Information Overload” and the “high recall” results during searching.

The survey was consisted of checking 3 pages of 17 multiple choice questions about the Internet search engine usages.

The generalities of the survey are the following:
Survey Target group: the first year students of University of Lugano, major in Science of Communication.
When: December 2003
Collecting data number: 68 questionnaires from circa 100 presences during the “Informatics I” class of Prof. Ghezzi.
Distribution of gender: M: 23 F: 45
The survey was conducted anonymously
Students had the possibility to answer each survey question also through more answers.

9.2 Relevant results

From the data analyzed, I can conclude that the Information Overload and the high recall level of search results are experienced by the users during the research.
1) Especially from the question “What is the major problem, you encounter with search engines?”, the answers were the following:

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect results</td>
<td>34</td>
</tr>
<tr>
<td>Too much information</td>
<td>20</td>
</tr>
<tr>
<td>Query formulation problems</td>
<td>19</td>
</tr>
<tr>
<td>Complexity of advanced functions</td>
<td>4</td>
</tr>
</tbody>
</table>

It is clear that there is the “Low precision High recall” problem in the search results. And there are the “**Information Overload**” and the Querying difficulties by users from which it derives the “**uncertainty**” to establish for communicating with search engines.
2) From the other question “If do you have an experience of getting lost during your search activity, why do you lose yourself?”, the answers were:

<table>
<thead>
<tr>
<th>Motive</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading all information from the results</td>
<td>29</td>
</tr>
<tr>
<td>Getting lost myself in browsing</td>
<td>16</td>
</tr>
<tr>
<td>Too much links opened at the same time</td>
<td>15</td>
</tr>
</tbody>
</table>

This is in line with the Information Overload problem, too. These three cases are the correlated situations between them, often. When an Information Overload situation occurs there are poor decision-making, difficulties in memorizing and remembering, and reduced attention span (Shenk, 1997).

![Figure 9-2  Motives of getting lost in Internet](image-url)
3) Another question was “Until where of the search engine results, do you look at?” , the answers were:

<table>
<thead>
<tr>
<th>Answer</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>41 students (61%) read the search results until they find out answer.</td>
<td></td>
</tr>
<tr>
<td>20 students (29%) read the first search results’ page.</td>
<td></td>
</tr>
<tr>
<td>4 students (6%) examine the first 5 links.</td>
<td></td>
</tr>
<tr>
<td>3 students (4%) examine search results until at the end of search results’ page.</td>
<td></td>
</tr>
</tbody>
</table>

It is possible to say that the majority of the users continued to look for accurate answer. They examine the relevancy of search results until they were convinced to consider what the right answer toward a query is.

Figure 9-3  Persistency to look for results
9.3 Some other results

1) How many times do you use search engines?
This graph shows that search engines are used frequently as I expected.

![Graph showing search engine use frequency](image)

**Figure 9-4 Search engine use frequency**

2) What is your major usage of Internet?
Also this graph shows that Information searching is one of the main use of Internet.

![Graph showing major usage of Internet](image)

**Figure 9-5 Major usage of Internet**
3) How do you get to web site?
This graph has one interesting result which is no one accesses Internet using directories. From this I can assume that students try to get to a goal easily without a par course fatigue.

![Figure 9-6  Internet navigation access modes](image)

4) Which search engine do you use generally?
This graph shows the predominance of Google.

![Figure 9-7  Major search engine used](image)
5) When you enter a query into the query form of a search engine, do you need to reformulate the queries? This result shows that students’ query reformulating tentative frequency is quite limited.

Figure 9-8  Frequency of query reformulation

6) Do you need to refine your query? The almost 60% of students have necessity to refining queries. From this results, unfortunately, the reason of un-refining does not show.

Figure 9-9  Query refining frequency
7) From the query refining activity which of them do you use?
This results show that major search engine user’s are aware of query refining function from search engine.

![Query refining used function](image)

8) How many different search engines do you use in the case that you cannot find information?
This graph shows that circa 70 % of students maintain in the same search engine, also 9% of students changing continuously. This circa 10 % students shows the uncertainty and the information seeking tendency.

![Search engine use type](image)

Figure 9-10  Query refining used function

Figure 9-11  Search engine use type
9) Would you store your right search keyword to your bookmarks on your browser?
This results shows that there are two types of community for organizing own search activity relationships.

![Figure 9-12 Keyword saving at bookmark](image)

10) What do you do when you find out good information from Internet?
This graph illustrates that the heterogeneous distribution of useful information storage.

![Figure 9-13 Useful link management modes](image)
9.4 Limits of survey

This survey was conducted by several limits:

1. I preferred to conduct the survey through the face-to-face communication. Since, by the personal experience which I rarely answered the e-mail based survey. After that, I had a conclusion that students would preferred to answer through the paper survey.

2. I prepared the multiple choice answers and one open answer to give own opinion. While the “helped” answers could give more immediately than open answers.
PART V – WORK IN PROGRESS

“We don’t know who we are unless we know where we’ve been.”

- Folk wisdom

10 New Approaches towards High Precision

10.1 Semantic web

A visual presentation of significantly related words.

- Semantic web

Berners-Lee, T. et al (2001) define "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

The current Web represents information using natural language and other formats as graphics, multimedia, page layout. However there are a couple of barriers to communication between those taking parts into the communication process. One of them is a language barrier. Not all online population uses the same language for communication. The second one is the semiotic problem; it is always present in each kind of communication.
In this specific case, the computer programs, even if they were so intelligent, they cannot understand human natural language as humans do. The programmers are human, but the major quantity of processing is done by machines. The automatic procedure is involved to solve it by great speed and precision. The Internet will reach its full potential when it becomes an environment where data can be shared and processed by automated tools as well as by people.

The raw information that may be on the Web is not in a machine-usable form. A person is still necessary to discern the meaning of the information and its relevance to our needs.

The Semantic Web addresses this problem in two ways. First, it will enable communities to expose their data so that a program doesn’t have to strip the formatting, pictures and ads from a Web page to guess at the relevant bits of information.

Secondly, it will allow people to write (or generate) files which explain -to a machine - the relationship between different sets of data. For example, one will be able to make a 'semantic link' between a database with a 'zip-code' column and a form with a 'zip' field that they actually mean the same thing. This will allow machines to follow links and facilitate the integration of data from many different sources.

This notion of being able to semantically link various resources (documents, images, people, concepts, etc) is important. With this we can begin to move from the current Web of simple hyperlinks to a more expressive semantically rich Web, a Web where we can incrementally add meaning and express a whole new set of relationships (hasLocation, worksFor, isAuthorOf, hasSubjectOf, dependsOn, etc) among resources, making explicit the particular contextual relationships that are implicit in the current Web. This will open new doors for effective information integration, management and automated services.20

20 http://www.ercim.org/publication/Ercim_News/enw51/berners-lee.html
10.2 Personalization

*Know your enemy and know yourself; in a hundred battles, you will never be defeated.*

-Sun Tzu, The Art of War

The ideal search engine would resemble the computer always responding to queries for information that precisely matched the needs of the user.

At the moment, search engines are far from that model. How can we know and recall diversified user’s need? First it is necessary to know the user; second, knowing means knowing some habitual patterns from them also from inference of other similar user pattern studies. Practically, it is based upon the inference from similarity of patterns. If that works it is possible to group or cluster. Reliability is improved if the site becomes more valuable to the user at each visit. After several visits, the site will fit the user’s needs so well that the user will be hesitant to move to another site that cannot offer the same service.

The team of Pennsylvania State University researchers looked at 60 sites listed on the SearchEngines.com web site. They asked three questions:

- How many search engine web sites offer personalization features?
- What features can be personalized?
- How accessible are the personalization features?
Only 13% of the 60 sites examined offered some level of personalization. And most of those personalization features were related to email, business and financial information, searching of a reference tool, such as yellow pages, entertainment listings, sports, and news headlines.

Why the lack of personalization features? The authors note that though personalization seems to be a simple concept, it’s difficult to implement. By its very nature, personalization means different things to different people, and there’s no one size that fits all solutions. Accessibility of features is another important consideration. In many cases, personalization features are available, but they are buried deep within a site and difficult for users to locate.

Another potential roadblock is the willingness of users to reveal information about them to fine-tune personalization features. Privacy policy statements are often confusing, jargon-rich screeds that arouse suspicion in the minds of users, leading to reluctance to share personal information.

User’s Searching Pattern Study

The most traditional way to achieve this goal is to encourage the user to enter a user profile: generally people act within a pattern, a habitual act. The Internet user acts also more or less following the same pattern. Analyzing the user’s profile is good information to individuate the user’s certain pattern. The downside of these common user profiles is that they require a large amount of up-front work by the user before any recompense can be gained. Therefore, users will only establish user profiles on sites that have already gathered their reliability.
Personalization works best if the site can gather information about users unobtrusively. For example, Amazon.com asks users to enter a complete shipping address for their first order, but if they order again, they will discover that the server has saved their address, making their second transaction much faster. Without any overhead, users enjoy faster processing if they keep ordering from Amazon.com than if they try out new online bookstores.

As another example of unobtrusive personalization, search engines could build up databases of user preferences by tracking which of the search result links a user actually clicked on. After collecting this data for a while, the search engine would know what kinds of links the user preferred and it could give such links higher ratings in subsequent searches.

The personalization of search tools entails matching results to user profiles. For example, the keyword "cook" might place nuovelle cuisine at the top of search results for someone who had recently visited nuovelle cuisine web sites, but might lead off with Web sites about a Family name of a professor. If you search using the keyword "surf," for example, the computer has no way of knowing whether you are looking for a windsurfing, information about how to surf safely, or simply a place, person, business or some other entity linked with "surf", it may as well be there the user intended it in the internet context. The problem is that the word "surf" has no meaning, or semantic content, to the computer.

Many search engines already use some rudimentary personalization features. AltaVista, Google etc most search engines use so-called geo-tracking technology to detect visitors' Internet protocol (IP) addresses and guess their geographical location. That can provide useful context for some searches, for example, in returning soccer-related results for a query on
"football" from a user based in the United Kingdom. But unfortunately, it is quite a local or language barrier too. E.g. Switzerland’s one third of population consists of foreigners, which means that geo tracking might not be that helpful.

Yahoo is also implementing personalized search features. Among other things, the Web portal has taken advantage of its relationship with visitors to deliver more tailored answers in specific areas. For example, its yellow pages, weather information and "My Yahoo" sections all use members' zip codes or other personal data to deliver tailored information.

Search leader Google has also shown an interest in the area. Using data-mining techniques, pattern recognition and natural-language semantic analysis to improve search results.
PART VI – CONCLUSION

“Communication is a process in which participants create and share information with one another in order to reach a mutual understanding” that is the Rogers and Kincaid’s (1981) theory which is the milestone of my thesis. The communication’s interactivity supports the mutual understanding of participants. The similarity (Berlo, 1960) of participants facilitates the communication. The homophilous communication (Lazarsfeld & Merton, 1964) easily creates the credibility (Hovland, Janis and Kelley, 1953), too. This convergence goal gives us the ultimate intention and sense of communicating with others. Consequently communication becomes a pleasant interaction process and a positive goal-oriented human activity.

In computer mediated human communication, the mutual understanding of participants gears up when the computer plays the role of a reliable human homophilous agent. The Internet Search Engine acts as an Information search agent in Internet. However, the Internet search engine is not a homophilous agent to the humans, yet. It causes also too much information transmitted to user.

As Berlo (1960) defined, in communication, messages do not mean by themselves. In that sense, the Internet Search engine communicates to the user too many inaccurate messages than the precise meaning. This situation verifies that a Search engine user’s message, specifically a query, frequently does not mean equally to the search engine. It could be comprehended also that the search engine knows the query but it does not understand the query or the user’s intention. Often, I had a positive experience to meet an expert who understood easily my intention at the beginning and during our conversation.
Therefore, the search engine users principally complain on search engines about the Information Overload phenomenon, due to the “high recall” state on the search results. Muramatsu & Pratt (2001) explained that “when users’ mental models differ from the actual underlying system model, users often become frustrated and fail to accomplish their tasks”.

To reduce the user's Information Overload phenomenon, the Internet search engines’ coordinators have to study more accurately the user's intention behind the user’s query and to implement the personalization solution.

Actually, the efforts toward the personalized search engine are a self-evident tendency of homophilous communication.

Consequently, the search results would achieve a “High Precision” state, regarding the specific user’s query.

The mutual understanding is never perfect even face to face communication. In other words, this could be considered as a fact of matter, impossible to change and part of the nature of communication itself. Therefore, the communication is an ever-changing, expanding, and back-and-forth motion process (Dance, F.1967). These vital and active components are an instrument of the social interactions that are typical for human beings. As well as these interactions, also communication changes and is dynamic.

I could imagine that one day, the Internet user will have the homophilous (Lazarsfeld & Merton, 1964) and empathic (Berlo, 1960) search engine agent, if we continue to move towards the communication for the better mutual understanding.
Bibliography


Eveland, J.D. (1986) *Diffusion, Technology Transfer and Implications: Thinking and Talking about Change*, Knowledge, 8(2)


Wiwatersrand, Johannesburg, 1980 Published: Communicare, 1981:2 (1), 42-61, last visited, May 19, 2004 from Web site:
http://www.und.ac.za/und/ccms/socialsemiotics/semiotics_semiology.htm


Internet sources

Communication theory
http://www.legamedia.net/lx/result/match/31917869034943a11179e7062a/index.php

Hussein, Y. (1997) Ethos (Ethical Proofs), Georgia Institute of Technology, last visited May 19, 2004, from web site:  
http://www.lcc.gatech.edu/gallery/rhetoric/terms/ethos.html

Latin roots of Communication, Colorado University, Department of Communication, last visited May 19, 2004, from Web site:  
http://www.colorado.edu/communication/meta-discourses/Theory/latin.htm

Nitcavic, R. Fundamentals of Public Communication, Ball state University, last visited May 19, 2004 from web site:  
http://www.comm210.com/m1c1tpc3.htm

Perseus lookup tools, search results for ‘comm’ latin, last visited May 19, 2004, from Web site:  
http://www.perseus.tufts.edu/cgi-bin/vor?lang=la&formentry=0&lookup=comm%26%23x016B%3Bn%26%23x12d%3Bt%26%23x12d%3Bo

Scott J. (2001) Interpersonal Communication courses, University of Kentucky, last visited May 19, 2004, from web site:  
http://www.cultsock.ndirect.co.uk/MUHome/cshtml/index.html

**Information Theory**


http://www.math.uwaterloo.ca/~pnijjar/cs692/misinformation.html

Principia Cybernetica Web (2001) *definition of Information*, last visited May 19, 2004, from web site: 
http://pespmc1.vub.ac.be/ASC/INFORMATION.html

Sharma, N. *Data Information Knowledge wisdom hierarchy*, School of Information, University of Michigan, last visited May 19, 2004, from web site: 
http://www-personal.si.umich.edu/~nsharma/dikw_origin.htm

http://www.unknowncountry.com/news/?id=3294

http://highered.mcgraw-hill.com/sites/0072400773/student_view0/glossary.html
Wall Street Journal, “Information Overload”, last visited May 19, 2004, from web site:
http://online.wsj.com/article_email/0,,SB106842510025265600-H9jeoNmlaV2oJyoZH6HaqWAm4,00.html

Internet study
http://www.press.umich.edu/jep/07-01/bergman.html

http://academics.smcvt.edu/sburks/Definition_search_engine.htm


Galactic Network, last visited May 19, 2004 from web site:
http://web.mit.edu/invent/iow/cerf.html

Graph structure in the web, www9 paper, IBM Almaden Research Center, last visited May 19, 2004, from web site:

Internet History, last visited May 19, 2004 from web site:
http://www.isoc.org/internet/history/brief.shtml
http://www.isoc.org/internet/history/brief.shtml#Timeline

Information pursuing pattern chart
http://www.gl.iit.edu/llc/webworkshop/outline.htm
Internet structure graphic, last visited May 19, 2004 from web site: 

Internet User statistics
http://cyberatlas.internet.com/big_picture/geographics/article/0,1323,5911_151151,00.html

*How much Information?* 2003, Regents of the University of California, 
last visited May 19, 2004 from web site:

J.C.R. Licklider homepage last visited May 19, 2004 from web site: 
http://medg.lcs.mit.edu/people/psz/Licklider.html

Koster, M. *The web robots FAQ*, last visited May 19, 2004, from web site:  
http://www.robotstxt.org/wc/faq.html

Lawson, J. (2002), *Google’s approach*, Internet tools for lawyers, last 
visited May 19, 2004, from web site:  
http://www.netlawtools.com/net_qa/net_qa_archives00-04.html

Link relationship graphic
http://www.touchgraph.com/TGGoogleBrowser.html

http://www.consumerwebwatch.org/news/searchengines/TOC.htm

Pagerank graphics
http://www.pcmag.com

Rafaeli, S. & Sudweeks, F. Networked Interactivity, last visited May 19, 2004, from web site:
http://www.ascusc.org/jcmc/vol2/issue4/rafaeli.sudweeks.html#Interactivity

Recall Precision graph, last visited May 19, 2004 from web site:
http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html

Search engine statistics
http://searchenginewatch.com/reports/article.php/2156481

Search results visualization tools
http://www.groxis.com

Search tools category (2003) Regents of the University of California, last visited May 19, 2004 from web site:
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/ThingsToKnow.html

The Pew Internet and American Life Project
http://www.pewinternet.org/reports/pdfs/PIP_Online_Pursuits_Final.PDF

Internet On line Dictionary

Search engines last visited May 20, 2004 from web site

Google, [http://www.google.com](http://www.google.com)
A9, [http://www.a9.com](http://www.a9.com)
Ask Jeeves, [http://www.ask.com](http://www.ask.com)
About, [http://about.com/](http://about.com/)
Overture, [http://www.overture.com](http://www.overture.com)

Meta-search engines
Vivisimo, [http://vivisimo.com](http://vivisimo.com)
Kartoo, [http://www.kartoo.com](http://www.kartoo.com)
Dogpile, [http://www.dogpile.com](http://www.dogpile.com)
Search, [http://www.search.com](http://www.search.com)
Mooter, [http://www.mooter.com](http://www.mooter.com)
Hotbot, [http://www.hotbot.com](http://www.hotbot.com)

Portals
Yahoo, [http://www.yahoo.com](http://www.yahoo.com)
Aol, [http://www.aol.com](http://www.aol.com)

News on Search engines
Annexes

Questionnaire

December 2003

Questionario per una ricerca della memoria

“Verso isole intelligenti navigando nell’oceano dell’informazione”

Doojin SONG, Scienze della comunicazione

Grazie mille per la sua gentile disponibilità!
Le risposte sono elaborate in forma anonima.

Per tutte le domande (tranne per quelle specificate) hai la possibilità di dare più risposte.

1. Quale è il tuo principale uso di Internet?
   a. Email
   b. Ricerca dell’informazione
   c. Chat(e.g. Chatroom, Messenger, ecc)
   d. Scaricamento delle file condivise
   e. News
   f. Altri funzioni:

2. Come cerchi le informazioni su Internet?
   a. Tramite i motori di ricerca
   b. Vado direttamente sul sito che mi serve, digito l’URL esatto
   c. “Navigando” a caso
   d. Chiedo prima delle informazioni a qualcuno, poi “navigo”
   e. Directories
   f. Altri metodi:

3. Se usi il motore di ricerca, quale tra quelli elencati utilizzi con maggior frequenza?(solo una risposta)
   a. Alltheweb
   b. Kartoo
   c. Google
   d. Overture
   e. Altavista
   f. Yahoo
   g. Altro:
4. I risultati dati dal motore di ricerca, sono inerenti alle informazioni che ti servono?
   a. No, quasi mai
   b. Sì, ma spesso devo riformulare le domande
   c. sempre ci riesco al primo colpo

5. Per formulare o riformulare le domande, usi certe funzioni ulteriori rispetto delle semplici parole chiave (key word)?
   a. No
   b. Sì, uso spesso le espressioni booleane
   c. Uso “ “ per trovare l’informazione che cerco
   d. uso spesso la funzione “Ricerche avanzate”
   e. altri metodi

6. Quando non trovi immediatamente le informazioni che cerchi e devi ripetere la domanda, che cosa fai?
   a. Uso lo stesso motore di ricerca
   b. Uso due motori di ricerca contemporaneamente
   c. Finché non trovo l’informazione che voglio, cambio di continuo il motore di ricerca.

7. Quando hai trovato le parole chiave che ti hanno permesso di trovare il corretto risultato, salvi la parola chiave e la pagina nei “preferiti”?
   a. Sì, mi potrebbe servire in un secondo momento
   b. No, Non vedo utilità
   c. No, perché:

8. Quando hai trovato un link giusto, qual è il tuo passo successivo?
   a. Salvo il link nel mio disco (duro o mobile)
   b. Salvo solo il link nei “Preferiti” del browser
   c. Stampo le pagine
   d. Stampo le pagine e salvo il link nei “Preferiti”
   e. Guardo e basta
   f. Altro:

9. I risultati che trovi mediamente i motori di ricerca, spesso sono eccessivi. Fin dove va la tua ricerca?
   a. Guardo solo i primi 5 risultati della prima pagina
   b. Guardo unicamente i primi 10 link, che corrispondono alla prima pagina
   c. Finché non trovo i link che mi servono, proseguo a guardare le pagine
   d. Proseguo fino alla fine delle pagine trovate
   e. Altro:
10. **Con quale frequenza usi i motori di ricerca?**
   a. Ogni giorno, almeno una volta
   b. Ogni settima, almeno una volta
   c. Ogni settimana, più di 3 volte
   d. Ogni tanto
   e. Altro:

11. **C'è un aspetto frustrante nei motori di ricerca?**
   No
   Si:
   a. Non so come formulare le mie domande
   b. Ci sono troppi link contemporaneamente
   c. Le informazioni sono spesso sbagliate (ovvero non corrispondono alle domande poste)
   d. Le funzioni aggiuntive sono complesse (e.g. ricerche avanzate)
   e. Altro:

12. **Se ti hai già capitato di “perderti” nella ricerca delle informazioni, sai spiegare il motivo?**
   a. Ho creato troppi link contemporaneamente
   b. Non sapevo come muovermi durante la ricerca
   c. Ho perso molto tempo nel leggere tutte le informazioni trovate
   d. Altro:

13. **Quando svolgi una ricerca (es. per la tesi, per un paper o per una presentazione), come cerchi le informazioni?**
   a. Dal generale al particolare
   b. Dal particolare al generale
   c. Dipende dall'argomento
   d. Altro:

14. **Hai altre osservazioni sulla ricerca delle informazioni tramite i motori di ricerca?**
   No
   Si:
Altre domande:

15. **Usi un’agenda cartacea o elettronica?**

16. **Ti è utile organizzare la tua giornata o la settimana?**

17. **Ti piace fare i viaggi non organizzati?**

Informazione generale

<table>
<thead>
<tr>
<th>Sesso;</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
</table>