Context based semantics for multimodal retrieval

Alexandra Dumitrescu Simone Santini

Escuela Politécnica Superior, Universidad Autónoma de Madrid
Madrid, Spain

ABSTRACT

This paper proposes an alternative to formal annotation for the representation of semantics, and presents an extension to it capable of handling multimedia (text and images) documents. The article argues that meaning is not a property of a document, but an outcome of a contextualized and situated process of interpretation. The consequence of this position is that one should not quite try to represent the meaning of a document (the way formal annotation does), but the context of the activity of which search is part.

We present some general considerations on the representation and use of the context, and a simple example of a technique to encode the context represented by the documents collected in the computer in which one is working, and to use them to direct search. We show preliminary results showing that even this rather simpleminded context representation can lead to considerable improvements with respect to commercial search engines both for text and images.

1. INTRODUCTION

In June 2006, Magen David Adom, the Israeli first aid society, officially joined the international red cross movement. In order for this to happen, one of the obstacles that had to be overcome was graphic symbolism. Up to that moment, the movement had used two symbols: the red cross, officially adopted in 1863, and the red crescent, created in Turkey in 1877, but adopted officially by the movement in 1929. The Israeli objected to the use of these symbols on cultural and religious grounds. The Magen David Adom had used the star of David as a symbol, but the Red Cross rejected the Israeli request to admit it internationally, since it is a religious symbol (the official explanation of the movement as to why the red cross is *not* a religious symbol is however a little thin). The impasse was resolved with the creation of the *red crystal* that Magen David Adom uses outside of Israel, although the appearance of the star of David is admitted alongside it.

If you are a follower of current theories of computational multimedia semantics, all this diatribe will make absolutely no sense to you. The current theories, in fact, will tell you that the meaning of an image, of a document, or of anything else, depends exclusively on its *contents* and, based on that, one can't imagine why the Israelis would object to the use of two lines, one horizontal, the other vertical, while would accept the use of four lines placed in the shape of a rhombus.

The problem, of course, is not with the Israelis, but with the the essentialist view of meaning held by computational semanticians, one that is is, if not unbearably naïve, at least desperately inadequate. Images are a node in a complex network of signification that goes beyond their content and includes other forms of textuality that go around them, as well as the cultural practices of the community that creates or receives images. In all lilelihood, he red cross wouldn't be the symbol of anything were it not for 2000 years of Christian tradition, and the Israeli would not object to it were it not for the religious symbolism of the image. There is, in other words, no meaning in images (or in anything else, for that matter) independent of the process of interpretation, a process that always takes place in a given context and as part of a given human activity. As the Italiaj philosopher Gianni Vattimo would put it: *there are no facts, only interpretations, and this too is an interpretation.*[11]

These considerations extend to the relation between images and words. There isn't *one* relation between text and images but, rather, a multitude of modalities: text can be used to explain images, images to explain text, images can set the mood in which text should be read, text and images can be two independent examples of the same category, text can contrast or contradict images, and so on. In this case as well, the most important thing that we should consider to make sense of the juxtaposition of text and images are the context of the search and the discursive practices of the environment in which the *synoptic text* (the organized layout of text and images) was created.

From the point of view of the computing professional, the essentialist view has the advantage of making annotation* possible or, at least, of answering the obvious question: "what do we annotate in an image?" If you are a naïve essentialist, the answer is obvious: since the meaning of an image is given by the object it contains, all you need to do is to write down (in a suitably formal language) what objects the image contains and, from that, you can easily *calculate* its meaning. With a certain flair for simplification, one could say that this point of view endorses statements such as "the image of a pencil means 'pencil'", or "the image of a nail means 'nail'". If somebody truly believes this, we see no better way to dispel it than reporting an example from a professional: the Mexican photographer Pedro Mayer†. In response to the question "would not a photograph of a pencil on a table always be just that, a photograph of a pencil on a table?" he replied:

> [Researchers in Peru had] the idea of using cameras to discover the codes being used by [poor Peruvian children]. They would ask a simple question and then elicit from the children [...] a response with a picture made with very simple cameras.
>
> They wanted to know what these children thought of "exploitation" [...]
>
> One child came back with a picture of a nail on a naked wall. At first the instructors thought that the child had misunderstood the idea of the experiment. But upon further investigation they found out that these children were living in an extremely poor town, several miles outside of Lima, and in order to make a little bit of money they walked every day all those many miles into town to shine shoes. And so that they did not have to carry back and forth the heavy load of the shoe boxes they rented a nail on the wall at someone's place in town. And the man that rented the nail charged them for this half of what they earned for the whole day.
>
> As you can see, sometimes a nail on the wall means much more than a nail on the wall. Or for a Cuban child the picture of a pencil on a table might have implications dealing with the blockade, as they had no pencils for a long time.

This example is interesting because, all in all, the contents of the image are quite irrelevant to their meaning. Better yet: the contents of the image are relevant only inasmuch they are apparently unrelated to their meaning, and assume a relevance once they are transformed by their intended meaning.

Phenomena like these are very common in all images manufactured to convey a message. Take a painting like, say, *Les demoiselles d'Avignon*, which represents four prostitutes. Representations of prostitutes abound in the history of art, from Greek pottery to Toulouse-Lautrec. What makes Picasso's painting meaningful is not the objects it depicts, but the way in which they are depicted, a way that inaugurates cubism. Documentary photography is no different. A picture that appeared in the French magazine *Paris match*[6] documented the first visit of Pope John Paul II to his native Poland. In the picture, the Pope doesn't appear at all: what we see is a forest of raising hands that salutes the arrival of an helicopter. It is left to the context, and to our knowledge of the conventions of news photography, to understand that the Pope is indeed in the helicopter.

This is an important reversal of the naïve assumptions about meaning: it is not the contents of the image that determines its meaning; it is the meaning that grabs the contents, repossesses them, and uses for its own semantic purposes. If we want to discuss an image, we should not ask ourselves what it contains but, rather, what instruments of communication have been used for its production, what relation between objects and states of the world have been singled out by the choice of this particular content, and what discursive practices make this relation an acceptable way of communicating. In the case of our first example case, a relation of metonymy leads from an object (a nail on somebody's wall) to the exploitative relation of which it is a part. The relation between the image and its meaning is given here by a particular context: that of its production. We can't really understand the image unless we are told the story of its production and of the context in which this happened.

[*] Some people prefer to use the charming but etymologically incorrect term *meta-data*.

[†] The example was reported by Fred Ritchin,[7] and we used it before[10]

*   *   *

These considerations (of a much wider generality than these simple examples) deny validity to what might be called the naïve theory of annotation, according to which describing the objects in an image and their relations reveal its meaning. One can still defend annotation, though: it is still possible, the argument would go, to fathom a system that would formalize not only the contents of an image, but also its relations with the textual, iconic, and iconographic elements that surround it, the discursive practices of its creation and (but here we are stretching plausibility) the complex relation between contents, context, and discursive practices on one side and meaning on the other. We believe that, even in this somewhat more sophisticated setting, the trust in annotation is misplaced for two orders of reasons: the dependence of meaning on the interpretation process, and the existence of that pesky inconvenient called human nature, which, as much as computing scientists are keen on ignoring it, keeps popping up whenever people are involved in the use of computers.

2. CONTEXT-BASED RETRIEVAL

In the light of the previous observations, it seems clear that one can't hope to simply encode the semantics of a document in manner independent of the hermeneutic act of reading: meaning is created anew with each interpretation, and is a result of that operation. Just like the tree falling in an uninhabited forest that makes no noise (although it does provoke acoustic waves), so a text, when it is not accessed, has no meaning (although it has the potential for signification). Our problems, then, are basically three: given a data access situation, we must (i) find a suitable context in which the data access is situated, (ii) find ways to formalize this context, at least to a certain degree (we are, after all, computing scientist, and we can only work with what we can formalize), and (iii) find ways in which the context can interact with the data to generate meaning.

Let us start with a fairly general theoretical model. We have said that the context in which a document is interpreted is essential to determine its meaning, that is, that the context *changes the meaning* of a text. We can also see things going in the opposite direction: the function of the semantics of a text is to *change the context* of the reader. If you are interested in literature, the context in which you look at American literature will not be the same after reading *Moby Dick*; if you travel on a freeway, your context will no longer be the same after seeing a speed limit sign. A document that doesn't change the context in which you act is, by definition, meaningless. We can express this situation with the following expression:

$$C_1 \xrightarrow{\mu(t)} C_2$$

where C_1 and C_2 are the contexts of the reader before and after interpreting the text, t is the text, and $\mu(t)$ is its meaning.

This is, as we have said, a very generic model, but we can use it to start answering some questions. For one thing, *is it possible to formalize meaning?* The answer of our model is that it is possible only to the extent that it is possible to formalize context. If C_1 and C_2 are formally defined in mathematical terms, then, and only then, it will be possible to give a formal definition of the function $\mu(t)$.

The properties of the "space of contexts" depend crucially on the properties of the representation of the context that we have chosen, and it is therefore difficult to say something more about meaning is we don't impose some additional restriction. A reasonable one seems to be that we be capable of measuring the degree by which two contexts differ by means of an operation $\Delta(C_1, C_2) \geq 0$ such that, for each context C, it is $\Delta(C, C) = 0$. We don't require, for the time being, that Δ be a distance. Now the meaning of a document d in a context C can be defined as the difference that d causes to C:

$$\mu_C(d) = \Delta(\mu(d)(C), C) \tag{1}$$

Within this theoretical framework we can analyze, at least in the first approximation, various existing approaches, and devise ways to extend them. In this general scheme, the ontological approach to meaning can be synthesized as a constant function:

$$\perp \xrightarrow{\mu(d)} C \tag{2}$$

that is, ontology assigns a meaning to a document independently of the context in which the document is interpreted. This fact results, in our model, in the creation of a constant context, which depends only on the document and not on what was there before.

The model that we have outlined in the previous section entails the demise of search and querying as identifiable and independent activities. The "death of the query" is the price that we have to pay for semantics, for if semantics can only be present in the context of a certain activity, then search can only be conceived as part of that activity, possibly as something of a very different nature for each different activity. In this analysis of the transformation of querying we receive some help from the Wittgensteinian notion of *Sprachspiel* (language game). Wittgenstein purposely didn't define exactly what a Sprachspiel was, on the ground that the different games are not related by a fixed set of criteria but by a "family resemblance".[12] We can say, with a certain degree of approximation typical of a formalizing discipline like computing science, that a Sprachspiel is a linguistic activity coördinated, at least partially, by a number of norms (some of which are implicit) that determine which language acts (the *moves* of the game) are permissible and which are their effects on the context of the game.

From this vantage point, what used to be called a query is not an activity but a type of move in a computing-linguistic game. The Sprachspiel, and therefore the type of move that the query represents, depends on the specific activity that we are considering. There are, on the other hand, certain common characteristics of the query move that, at least pragmatically, we can take as a unifying trait. For starters, we will in general consider a specific and relatively restricted class of activities: those that take place primarily with the aid of a computer. In the course of these activities, we manipulate data of the most disparate kinds, from tables of numbers to geometric models, images, and so on. This manipulation process, inasmuch as it is linguistic (and the terms linguistic should be taken here in a very broad sense) can be seen as a Sprachspiel; the data on which one is working and those conceptually related to them constitute the context of the game.

Sprachspiel are connected solely by a tenuous "family resemblance" and just because some members of the family can usefully be represented as diagrams, this doesn't imply that they all can. So, for the moment, we will just assume that:

i) there are a number of activities that are carried out on a computer and whose context can be partially identified with the contents of the computer on which they take place;

ii) these activities may be modeled as games in which certain moves (e.g. writing a paragraph) are available to operate on the context; the purpose of the game is to reach a context with certain desirable characteristics (e.g. a context in which a satisfactorily compiled tax form is present);

iii) some of these moves entail introducing data in the context without actually producing them; we call these moves *queries*.

3. IMPLEMENTING CONTEXT

The practical problems posed by the general orientation presented here include how to capture ongoing activities, how to represent them and, to the extent that it's possible, formalize them, in such a way that they can be used as a basis for data access. In general, of course, this is impossible. If a person is, say, shopping for detergent and wants to search the internet for brands with certain characteristics, there is very little hope that we can represent the activity "shopping for detergent" in a computer system: we are in this case in the presence of a physical activity that leaves no *digital trace*, so to speak.

On the other hand, a significant number of daily activities are, for many of us, executed on or with the aid of a computer, and they do have a digital trace, one that can be recorded and used as a context for a language game carried out as part of that activity. Suppose that we are preparing a presentation for a conference to which we had submitted a paper and that, during this process, we need to clarify a point or to look for an illustration for the presentation. In order to prepare the presentation, we have created a document in a directory (let us say the directory *presentation*) where we have possibly copied some documents that we thought might be useful. This directory is likely to be placed in a hierarchy as in figure 1. Its sibling directories will contain documents somehow related to the topic at hand although, probably, not so directly as those that can be found in the

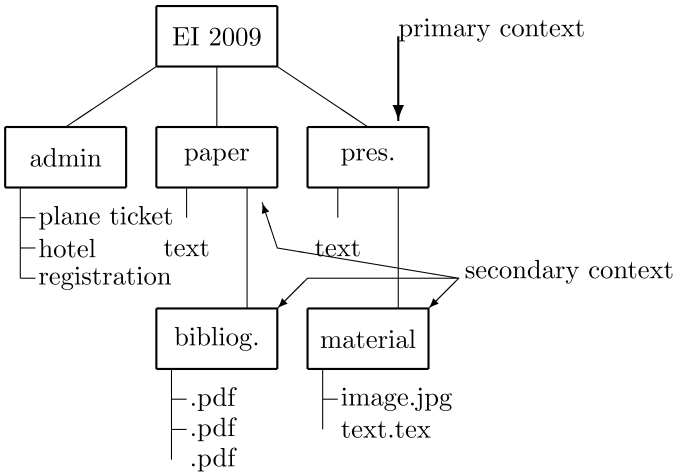

Figure 1. The structure of directories and context for the preparation of a presentation.

work directory. The siblings of the conference directory (and their descendants) will contain documents related to my general area of activity, although not necessarily directly related to the topic of the presentation. This information, suitably encoded, will constitute the context of the game. In order to create and play it, we have to specify two things: how to represent the context and how the various moves that the game allows will modify it; in particular, in this example, how the query moves of the game modify it.

3.1. Context representation

In order to build a representation, we consider two types of contexts: the *primary context* consists of the directory in which the current activity is taking place; the *accessory context* consists of any other directory that contains material in some capacity related to the current activity. The accessory context contains, in general, the descendants of the work directory and, possibly, its parent. This choice is somewhat *ad hoc*, and it is foreseeable that different systems will choose to use different *context policies* in order to determine the accessory context.

In each directory we create first a representation that takes into account only the files contained therein; we call such representation the *generator* of the directory. Then, for each directory, we create a further representation, called the *index*, built based on the generator of the directory itself (viz. of the primary context of the activities that take place there) and of the accessory contexts, as per the specific context policy adopted.

In the above example, in each of the six directories a generator will be created, with an appropriate representation of the context of that directory (that is to say, a representation of the documents that appear in the directory). The generators of the *pres* directory (the primary context) and of the directories *paper*, *bibliog.* and *material* (the accessory context), will join using appropriate operators, to form the index of the context of the search, which is stored in the directory *pres.* It must be noted that the construction of the index through generators supposes a hypothesis of compositionality of the context representation: the representation of the global context of two or more directories depends only on the representations of the local contexts and the relation between directories.

Let us begin by considering the construction of a generator, that is, of the context of a single directory that depends only on the documents found in the directory. In this example, we represent contexts using a technique similar to that of the semantic map WEBSOM.[3] This semantic map presents two features that are essential in our case: the representation of context by means of *self-organizing maps*[4] in the Euclidean space of words, and the use of *word contexts* as a working and learning unit of the map. Note that we are using the technique in a very different capacity than that for which it was originally conceived: we do not use it to represent the data space but the context; that is, its function is not indexing as in (Kaski, 1997),[3] but query formation.

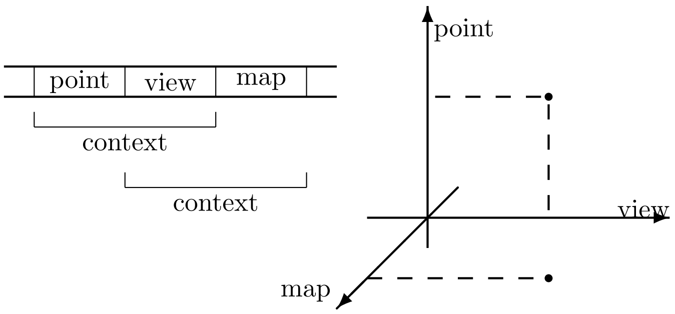

Figure 2. The geometry of the words context.

The *self-organizing* map forms a sort of non-linear *latent semantic*[1] space, and this non-linearity will is when making changes in the context (e.g. to express a query, as we shall see shortly).

Many representations of documents use the frequencies of words of the document; this representation is insufficient for our problem because if we use only a word by itself, the semantics that derives from the colocation of the words, namely the semantic component that is needed to solve problems like the polysemy, will be lost. On the other hand, in the technique that we will use, the fundamental unit of representation that is extracted from the document is not the word, but a group of words, that is called *word context.* The number of words of the *word context* may vary, in this work we consider the simplest case: two words, namely, we will consider pairs. Each pair of consecutive words in the text is seen as a symbol to which we assigns a weight proportional to the number of times the symbol (in other words, the pair of words) appears in the text (figure 2, left).

These pairs are represented in the typical geometric space of many information retrieval systems, a space in which each word is an axis. Since our basis are the contexts, the points in this space are not points in one of the axes (as in the case of simple words: each point is a word with its weight), but points in two-dimensional sub-spaces: each pair is a point in the plane represented by the two words that compose it. Using more complex contexts will result in points contained in spaces of higher dimension. As customary, before considering the words for the construction of indices, we will perform stop-word removal and stemming.

The *index* is a union of the generators of the primary and accessory contexts. In the case of our reference activity, the accessory context is composed of the descendants and the parent of the work directory. The weight of the pair constitute by the word number i and word number j (in other words, the word pair who has values in the e_i and e_j axes of the space of words), which may appear in several directories of the work context, is $omega^{ij}$. Each generator that we use in order to compute the context has its own weight for the pair, assigned depending on the frequency of that pair in the local directory. Let ω_P^{ij} be the weight for the pair i, j in the primary context folder, S_k be the kth directory that composes the accessory context ($k = 1, \ldots, S$), and ω_k^{ij} the weight in that directory. Then the weight of the pair i, j in the context, ω^{ij} is given by the weighted linear combination:

$$\omega^{ij} = \gamma\omega_P^{ij} + \frac{1-\gamma}{S}\sum_k \omega_k^{ij} \tag{3}$$

where γ is a constant, $0 \leq \gamma \leq 1$.

The map consists of a matrix of $N \times M$ neurons, each neuron being a vector in the word space;if the context is composed of T words, the neuron μ, ν ($1 \leq \mu \leq N$, $1 \leq \nu \leq M$) is a vector

$$[\mu\nu] = (u_{\mu\nu}^1, \ldots, u_{\mu\nu}^T) \tag{4}$$

The map learning is being developed under the stimulus of a set of points in input space, each point representing a pair of words *(word context).* Given a total number of P pairs, and given that pair number k consists of the words number i and j, the corresponding point in the input space is given by

$$p_k = (\underbrace{\overbrace{0, \dots, \omega^{ij}}^{i}, 0, \dots, \omega^{ij}}_{j}, 0, \dots, 0) \tag{5}$$

where ω^{ij} is the weight of the pair of words determined as in (3). During learning the p_k vectors are presented several times to the map. We call *event* the presentation of a vector p_k, and *iteration* the presentation of all vectors. Learning consists of several iterations. An event in which the vector p_k is presented entails the following operations:

i) Identify the "winning" neuron, in other words the neuron that is closer to the vector p_k:

$$[*] = \min_{[\mu\nu]} \sum_{j=1}^{T} (p_k^j u_{\mu\nu}^j)^2 \tag{6}$$

ii) The winning neuron, $[*]$, and a certain number of neurons in its "neighborhood" are moving toward the p_k point an amount that depends on the distance between the neuron and the winner one and the number of iterations that have been performed so far. For it, we define the *distance* between the neurons of the map as:

$$\|[\mu\nu] - [\mu'\nu']\| = |\mu - \mu'| + |\nu + \nu'|, \tag{7}$$

for $t = 0, 1, \dots$ the counter of the iterations of the learning. We define a function of environment $h(t, n)$ such that

$$\begin{array}{ll} \forall t, n \geq 0 & 0 \leq h(t, n) \leq 1, h(t, 0) = 1 \\ & h(t, n) \geq h(t, n + 1) \\ & h(t, n) \geq h(t + 1, n) \end{array} \tag{8}$$

and a coefficient of learning $\alpha(t)$ such that

$$\forall t \geq 0, 0 \leq \alpha(t) \leq 1, \alpha(t) \geq \alpha(t + 1) \tag{9}$$

Then each neuron $[\mu\nu]$ of the map moves toward the point p_k according to the learning equation

$$[\mu\nu] \leftarrow [\mu\nu] + \alpha(t) h(t, \|[*] - [\mu\nu]\|)(p_k - [\mu\nu]) \tag{10}$$

The function h generically corresponds to an environment of the winning neuron that is done smaller as it increases the number of iterations. In this work the environment function is the Gaussian $h(t, n) = \exp(-n^2/\sigma(t)^2)$, con $\sigma(t) \geq \sigma(t + 1) > 0$.

At the end of the learning process the map is laid out in the space of a word in a way that, in the extreme case of an infinite number of neurons that form a continuum, it optimally approximates the distribution of the points in the space.[9] This map represents the semantic space of the context and, as we mentioned in the previous section, can be assimilated to a nonlinear form of latent semantics.

3.2. Multimedia Context

The context representation that we have considered so far includes essentially text documents and, as we shall see, is used essentially to retrieve text document. One interesting question then is: how can we extend the context based retrieval approach to multimedia documents? That such an extension would be desirable stems from two observations:

i) as we have briefly shown in the first sections of this paper, the meaning of an image or of a video depends on much more than the contents of that image or video; it depends crucially on the reception context and on its interaction with the discursive practices that presided the creation of the artifact;

ii) even though the contents of an image have in many cases a fairly good correlation with its meaning (at least in the majority of interpretative contexts), it is well known that most of the features we can reliably extract from images are exceedingly poor at representing semantically relevant elements (the *semantic gap* problem[8]).

That is, in multimedia we can expect even more benefits from the use of context than we do in retrieval of textual documents. Multimedia context based retrieval can be seen in two different way, depending of whether one includes multimedia data in the representation of the context or not. A first, minimalist way of doing multimedia context based retrieval (the one about which we present some data in this paper) is based on the contextual extension of the *googl-ish*, so to speak, way of retrieving images. In many internet-based image search engines, in fact, retrieval is based not on the image data but on the text that surrounds the images: the user types a query, the query is matched with the name of the image, the text on the web page near to it, the text of links that point to the image, etc., and the images with the closest matching text are returned as results. The same context techniques that we use for text documents can be used to improve the retrieval of images using associated text. We will show that, even though we still don't have conclusive results, context techniques appear very promising in this area.

Another way of doing multimedia context based retrieval is to use multimedia data as part of the context. In general, the contents of a work folder will contain images as well as documents, and this creates additional information that we can use for context creation. The fact that we represent words in a vector space facilitates the inclusion of images in the contextual representation. Let us, first of all, assume that images are represented using n normalized features that compose a feature vector $\psi \in \mathbb{R}^n$. By adding n axes to the word space, we can represent images in the same space as words. We can assume that all the weights are normalized to $[0, 1]$, so that each portion of an image and each word context can be represented as a point in the unit cube of that space. A multimedia document, as well as a multimedia context, is then a cloud of points in the cube.

This model can be improved: we represent text using word contexts, that is, colocations of groups of words, but we represent images using feature points that make no reference to the text. In other words, this model has so far no way of representing the colocation of text and image fragments. Colocation works somehow differently for images than it does for words. In the case of words, the rules of syntax dictate that, with usable reliability, logical closeness is related to physical closeness, and this is why taking groups of contiguous words results in semantically meaningful associations. In the case of images, things are not that simple, and depend on the discursive practices of the domain from which images are derived. In a typical text and image layout, the text on the immediate side of the image may be only marginally related to its contents, since images are often placed at a certain distance from the portion of text they refer to, either because of graphical conventions or, in web pages, because of rendering variations between different browsers. However, in a newspaper article or in a mainly textual book, the text immediately below the image is a caption, and bears a direct relation with the contents of the image. But if the image is advertising, it has normally no relation with the text of the document—as a matter of fact, its relation with the document is that of a parasite towards its vehicle, and should not be considered as part of the document at all.

There are other forms of text that, although physically nowhere near the image in the rendered document, are nevertheless closely associated to it, such as the "ALT" text in the image tag or, if the image is also a link, the text it links to.

Be it as it may, we can assume the existence of a *logical distance* of sort between a portion of an image (as characterized by the feature vector ψ), and a portion of text in the same document. This distance (or rather a decreasing function of it) can be used to associate the features that describe regions of the image to a word context composed of the n words closest to it (figure 3).

3.3. The query

In its most complete and general form, the procedure of a query is composed of four phases:

i) through an appropriate user interface or with a program that the user is using, an initial specification of the query is collected, we will name it the *proto-query.* The proto-query can be formed by a few words typed

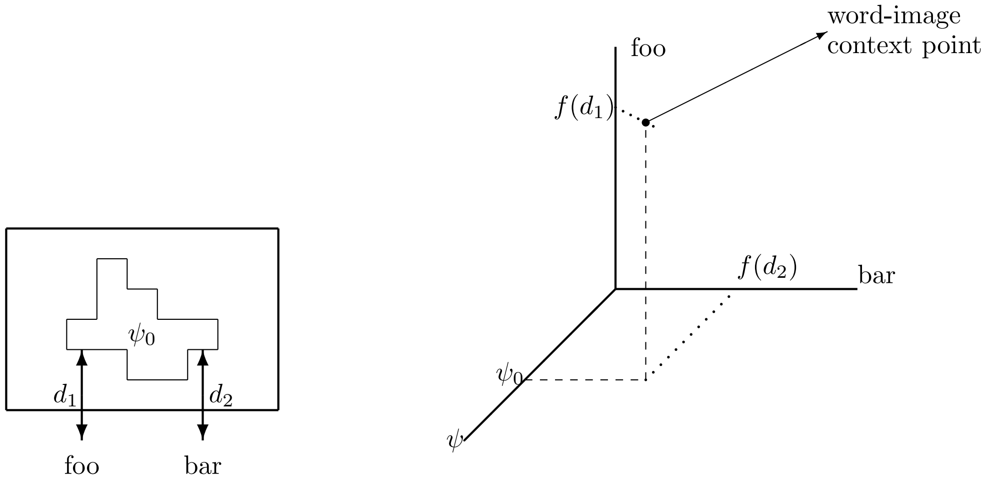

Figure 3. The geometry of the image-and-word context.

by the user, a paragraph that the user is editing, etc.. In a multimedia system the proto-query also contain an indication of the type of the document that's being searched (text, image, video, etc.)..

ii) The proto-query is used to change the current context, transforming it into a *objective context.* In practice, the configuration of the map (index) of the current directory is modified through a partial learning, which will give the context a *bias* towards the proto-query. The resulting configuration from this learning could be considered, in some way, as the interpretation of the proto-query in the actual context.

iii) The difference between the actual and objective context is the *differential context* and, in our model of semantics, corresponds to the semantic of the ideal document that is searched for: the document that is assimilated to the current context, will transform it into the objective context. An opportune codification of the ideal document is created and sent to the search server to retrieve the documents that more respond to that profile.

iv) The documents elected (e.g. read or downloaded) become part of the context: a new learning is run so that the current context reflects the new situation.

This general model of a query assumes the existence of a search service *(search engine)* capable of managing them. The construction of such a service is one of future goals of our work. For the moment, our objective is to demonstrate the role played by the context using it to focus searches on existing services. Therefore, it is necessary to transform the differential context into a list of words with weights, because the search services only accepts (if accepts) this type of queries. Obviously this type of query can not make an optimal use of the possibilities of context but, we repeat it, at this moment our goal is simply to evaluate the influence of the use of the context in the search. In our tests, the proto-query P is a set of keywords u_i. A keyword that correspond to the i word of the space is represented as the vector $e_i = (\overbrace{0,\ldots,1}^{i},0,\ldots,0)$. For simplicity we assume that every word in the query has the same weight w_i. Therefore, the query Q, formed by q words, will be represented as a point in the T-dimensional space: $Q = w\sum_{u_i \in P}^{q} e_i$

This vector is used for a partial learning process using the algorithm presented. During this process the neuron $[\mu\nu]$ is moved to the position $[\mu'\nu']$. The differential context is given by the differences of the neurons positions, $\delta_{\mu\nu} = [\mu'\nu'] - [\mu\nu]$ for each $[\mu\nu]$ in a neighborhood of the winning neuron (the closest neuron to the vector Q).

Projecting the vector $\delta_{\mu\nu}$ on the axes of the words, we get the weights of the words given by this neuron: $\delta_{\mu\nu} = (v^1_{\mu\nu}, \ldots, v^T_{\mu\nu})$. The *non-normalized weight* of the word i is given by the sum of their weights relative to all the neurons in a neighborhood A of the winning neuron

$$V^i = \sum_{[\mu\nu]\in A} v^i_{\mu\nu} \tag{11}$$

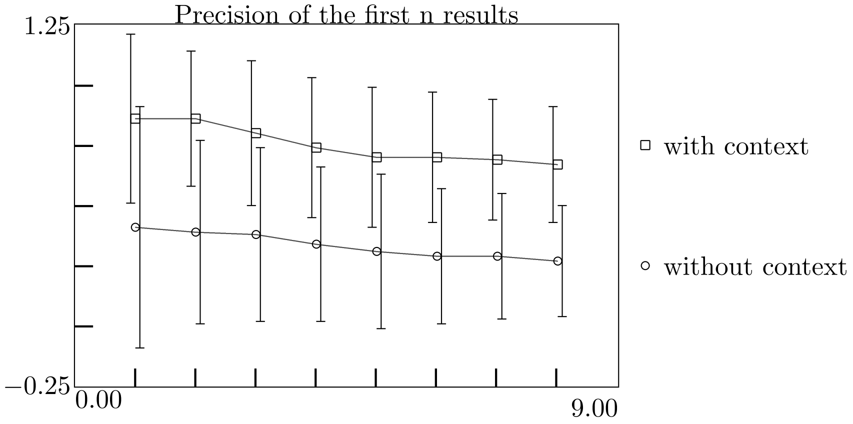

Figure 4. Precision of the results, with and without context.

Considering only the K words with greater weights, and normalizing the vector of weights for these words we obtain the query that will be send to the search engine, composed of a set of words each one associated with a weight.

* * *

Testing fully the context approach is quite problematic at this time for lack of a proper contextual server and its data base infrastructure. In order to obtain some preliminary indications, we used the limited weighting capabilities offered by the *google* commercial search engine (www.google.com). The contextual query was translated in a collection of weighted terms, and weighting was roughly approximated through positioning and repetition in the search engine query. As context, we considered, for the example reported here, the directory structure in the computer of one of us (Santini), and as working directory a directory with several columns by that author for the magazine *IEEE Computer*. We queried the search engine with 32 query terms, with and without the context, and measure the fraction of the first n documents that were considered relevant, for $1 \le n \le 8$. Given the generic and varied nature of the columns contained in the directory, a document was considered relevant if it was about computing. Note that the measure adopted here is the precision of the result. Not having a fixed corpus of documents in which we searched (or, rather, being the corpus the whole data base of the search engine) we couldn't measure recall. The results are shown in figure 4. It is evident even without a detailed analysis that the difference is large and statistically significant. Qualitatively, the difference depends on the particular query that is being made. Very technical words, whose semantic span is very limited to begin with, benefit little from the context, and fetch basically the same results with or without it. A query word such as "algorithm", for instance, is quite unlikely to fetch documents not related to computing, regardless of the presence of context. On the opposite side, queries with ambiguous terms, such as "sort" (data sort in computing, an approximation of qualities in the common language) gave the most dramatic improvements when context was used.

* * *

In order to explore the viability of context-based retrieval for images and other multimedia documents, we conducted some preliminary tests using the "minimalist" model previously described. The results reported here refer to two contexts: the first is the same computing context used in the previous test, whilst the second is a collection of documents used to prepare reports in a neurophysiology project. The queries were created in the same way as it was done for the text search, and sent to the *google image* search engine for image retrieval. The users of the system were instructed to consider an image as relevant if, in their judgment, it could conceivably be used as a technical illustration in a presentation or a paper in the subject of the context (computing or

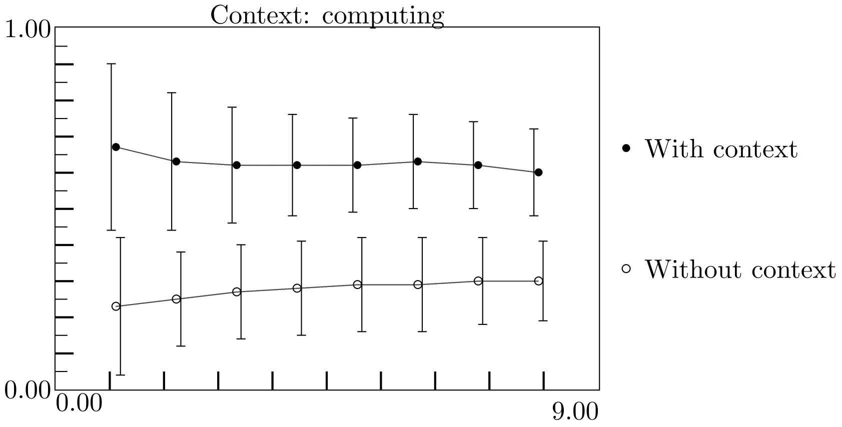

Figure 5. Precision of the results, with and without context.

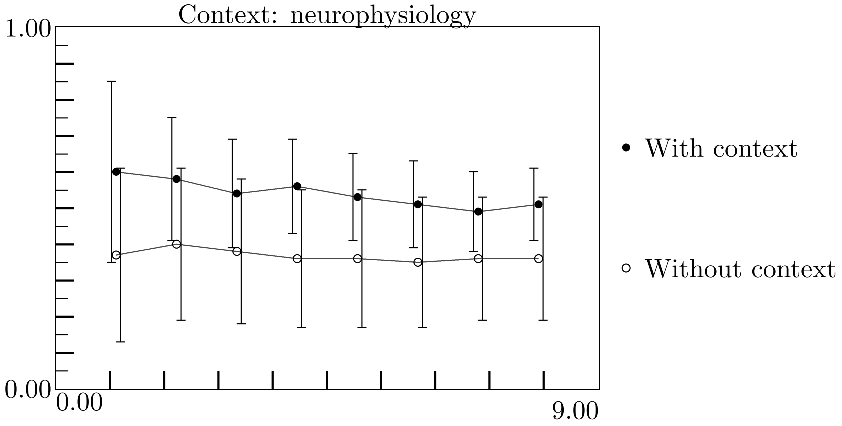

Figure 6. Precision of the results, with and without context.

neurophysiology). The results are reported in figures 5 and 6. In both cases, the use of context entailed a statistically significant improvement in precision but, from a superficial look at the data, the improvement appears more pronounced in the case of computing. The reasons for this might be several, some of which are probably due to the inherent characteristics of the search engine rather than to the use of context, but it is likely that, on the context side, the difference is due in good part to the different nature of the words used in the two contexts. Neurophysiologists, by and large, use neologisms with relatively uncommon Greek roots to indicate important concepts so, in this case, even a single word is sufficient to characterize the context with sufficient precision, and further contextual information brings only marginal advantages. Computing, on the other hand, tends to borrow words from other areas without modifying them, or making only superficial adjustments. That is, computing words are much more ambiguous than neurophysiological ones, and are by themselves, a poor characterizer of context. In this case, we can expect that a more complete characterization of context will bring the greatest advantages, as is indeed observed in the experiments.

It must ve stressed again that these are absolutely preliminary results that we use only as an indication of the viability of context use in multimedia. We do believe that a proper way of using context in this case entails the creation of multimedia contexts and the design of specialized context-sensitive search engines.

4. CONCLUSIONS

We have argued that formal annotation, and the general ontological programme that comes with it, might not be the proper way to consider the problem of the meaning of multimedia documents and, in general, to frame the issues related to multimedia semantics. This is not a majority opinion, not by a long shot, and there are a few reasons that contribute to its unpopularity and to the exclusivity of the attention given to annotation and ontology.

Apart from the attractiveness of a certain kind of naïve folks psychology, there is the understandable inertia of an established position on which a considerable intellectual and financial investment has been made. This phenomenon is quite well understood in the modern epistemological literature. One of the distinctive features of Kuhnian epistemology[2] is precisely taking into account such cultural and institutional phenomena.

There is also a point related to the economy of the commercial web (which, unlike ten years ago, today represents the vast majority of the web today). The model of meaning assumed by the semantic web is very appealing to web companies because, if meaning is inherent in a text, it can be owned, bought, and sold like other goods. Lyotard, in 1979, observed a similar phenomenon regarding knowledge: "knowledge is and will be produced in order to be sold, is and will be consumed in order to be valued in production: in both cases, in order to be exchanged"[‡]. Lyotard considers this phenomenon as a natural consequence of the computerization of knowledge: "[knowledge] can go through the new channels [(those of informatics)] and become operational only if it can be traduced in amount of information"[§]. It is not too daring, then, to expect that a similar change will occur with respect to meaning once this has been codified in formal annotations: only meaning that *can* be codified will survive, and this will do so only in order to be exchanged as merchandise.

This paper has presented the outline of a different model of meaning, one in which the reader's context plays a preponderant rôle. We have presented a simple framework in which we are currently experimenting with this model, a framework that in the future will be extended in different directions: on the one hand, the integration in this framework of more formal representations, at least for those parts of the context that can be formalized; on the other hand, the development of suitable data base techniques to make this kind of query efficient.

Our purpose will be, on the one hand, to build a context-based data access client (configured as a plug-in to some word processing or presentation program, if possible) to make context based retrieval on general web sites and repositories and, on the other hand, to build a context-based access server. The latter will be akin to the servers built for search engines such as yahoo or google but, while these servers do not coöperate with the user's computer (apart from the elementary communication necessary to retrieve the query and return the results), the server that we consider here will be integrated with the user's computer from which it will derive the current context, and with which it will coöperate to support interaction.

REFERENCES

1. Scott Deerwester, Susan T. Dumais, George W, Furnas, Thomas K, Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 2000.
2. Thomas Juhn. *The structure of scientific revolutions.* Chigago:Chicago University Press, 1996.
3. S. Kaski. Computationally efficient approximation of a probabilistic model for document representation in the WEBSOM full-text analysis method. *Neural Processing letters*, 5(2), 1997.
4. T. Kohonen. *Self-organizing maps.* Heidelberg, Berlin, New York:Springer-Verlag, 2001.
5. Jean-François Lyotard. *La condition postmodèrne.* Paris:Editions de minuit, 2001.
6. Le pape en pologne. (C) Paris Match, 1979.
7. Fred Ritchin. *In Our Own Image.* Aperture, 1999.
8. S. Santini. *Exploratory image databases: context-based retrieval.* San Diego:Academic Press, 2001.
9. Simone Santini. The self-organizing field. *IEEE Transactions on Neural Networks*, 7(6):1415–23, 1996.
10. Simone Santini. Image semantics without annotation. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web.* IOS Press, 2003.
11. Gianni Vattimo. *Oltre l'interpretazione; il significato dell'ermeneutica per la filosofia.* Laterza, 1994.
12. Ludwig Wittgenstein. *Philosophical Investigations.* Prentice Hall, 1973.

[‡] Lyotard,[5] p. 14, our translation.

[§] *ibid.* p. 13