Repositorio Institucional de la Universidad Autónoma de Madrid

https://repositorio.uam.es

Esta es la **versión de autor** de la comunicación de congreso publicada en:
This is an **author produced version** of a paper published in:

Semantic Multimedia: 4th International Conference on Semantic and Digital Media Technologies, SAMT 2009 Graz, Austria, December 2-4, 2009 Proceedings. Lecture Notes in Computer Science, Volumen 5887. Springer, 2009. 173-176.

DOI: http://dx.doi.org/10.1007/978-3-642-10543-2_19

Copyright: © 2009 Springer-Verlag

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

Incremental context creation and its effects on semantic query precision

Alexandra Dumitrescu and Simone Santini*

Escuela Politécnica Superior, Universidad Autónoma de Madrid

Abstract. We briefly describe the results of an experimental study on the incremental creation of context out of the results of targeted queries, and discuss the increase in retrieval precision that results from the incremental enrichment of context.

1 Introduction

This paper is, in essence, a progress report on an activity that was presented last year at this very conference series. Last year, we presented a conceptual model (and its practical incarnation) that used the documents in a person's computer to create a *context*, and used it to conduct semantic searches on the web. The theoretical bases for the semantic model that we use are radically different from the Tarskian semantics of the semantic web, and can be traced on the one hand to the hermeneutic tradition and, on the other hand, to the anglo-saxon philosophy of language that assumed the findings of the second Wittgenstein [3].

Last year we presented our model and evaluated its behavior using complete and fixed contexts. We took two rich collections of documents about computing and neurophysiology, respectively, and determined how much the presence of this context would improve the precision of the results of generic queries in these two areas. The results were quite encouraging, sometimes more than doubling the precision of the same query without context.

In this report, we briefly discuss an experimental study of context formation. We used the system iteratively to see how quickly, starting from a situation without any context information, we could build up a context that allowed the considerable improvements that we observed last year.

In order to make the paper reasonably self-contained, we will include a brief description of our context model. In [4], the context was based on a set of directories with a sub-directory relationship that caused some complication in the derivation of the model. Since the results that we describe in this paper are based on the contents of a single directory, we will described a simplified model that doesn't take into account structure, and that is sufficient for our present purposes. For details on the complete model, the reader is referred to [4].

* This work was supported in part by *Consejería de Educación, Comunidad Autónoma de Madrid*, under the grant CCG08-UAM/TIC/4303, *Búsqueda basada en contexto como alternativa semántica al modelo ontológico*. Simone Santini was in part supported by the *Ramón y Cajal* initiative of the *Ministero de educación y ciencia*. Alexandra Dumitrescu was in part supported by the *European Social Fund*, *Universidad Autónoma de Madrid*.

2 Context definition

Our starting point for the creation of context is a collection of documents. In the complete system, these are the documents contained in the working directory from which the user starts a query as well as the documents retrieved and downloaded in the course of previous queries. In this case, we are interested in studying the process of context formation, so we will not consider any document in the working directory, making the experimental assumption that the first query is made from an empty context without any document. The context is created by downloading query results and accumulating them. We wish to emphasize that this is *not* the way our system is meant to be used. In general, all the documents relevant to a given activity are used to make up the context. The choice of the particular operational mode used here is in accord to the experimental design that we present in this paper.

We use two context representations, the second being built based on the first. The first representation, which in the general model we call the *syntagma*, is syntactic, and in the present incarnation of the system is a *point cloud* representation. The second, called the *seme* is a semantic representation implemented, in this case, as a self-organizing map that constitutes a *latent semantic manifold*, that is, a non-linear low-dimensional subspace of the word space that capture important semantic regularities among words. The technique is based on the self-organizing map WEBSOM [1], but while WEBSOM and other latent semantic techniques have been used so far mainly for the reprsentation of data bases, we shall use them as a context representation.

Consider that we have already executed a number of queries and collected certain documents that are considered to be relevant. We join all these documents in a single large document and apply standard algorithms for stopword removal and stemming. The result is a series of *stems* of significant words, from which we consider pairs of consecutive stems (*word pairs*).

Let the words (terms) of the document $[t_1, \ldots, t_W]$, (ij) be the pair formed by the word t_i followed by t_j, P the set of all pairs found in the documents, and N_{ij} the number of times that the pair (ij) appears. The pair (ij) is given a weight $w_{ij} = \frac{N_{ij}}{\sum_{(hk)\in P} N_{hk}}$ The pairs are represented in a vector space whose axes are the words $t_1, \ldots, t_W$. In this space, the pair (ij), with weight w_{ij}^D is represented by the point $p_{ij} = (\overbrace{0, \ldots, w_{ij}^D}^{i}, 0, \ldots, \underbrace{w_{ij}^D}_{j}, 0, \ldots, 0)$. All these points form the point cloud I_D.

The point cloud thus built is used as the training data for a self-organizing map deployed in the term space. The map is a grid of elements called *neurons*, each one of which is a point in the word space and is identified by two integer indices: $[\mu\nu] = (u_1^{\mu\nu}, \ldots, u_T^{\mu\nu}) \quad 1 \leq \mu \leq N, 1 \leq \nu \leq M$ The map is discrete, two dimensional with the 4-neighborhood topology. On it we define a *neighborhood function*, $h(t, n)$, which depends on two parameters $t, n \in \mathbb{N}$; n is the graph distance between the neuron whose neighborhood we are determining and another neuron, t is a time parameter that increases as learning proceeds. The function decreases with the distance from the given neuron and "shrinks" with time. We also define a *learning parameter* $\alpha(t)$, decreasing with time.

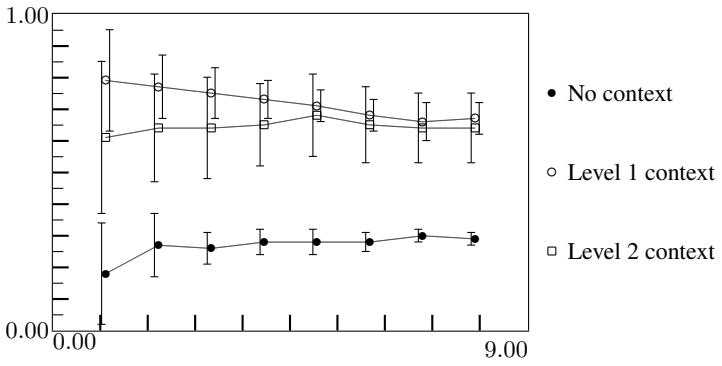

Fig. 1. Precision of the results for the computing context.

All the points in I_D are presented to the map. We call the presentation of a point $p \in I_D$ an *event*, and the presentation of all the points of I_D an *epoch*. Learning consists of a number of epochs, counted by a counter t. The neurons of the map are at first spread randomly in the word space; then, for each event consisting of the presentation of the point p, we apply the standard Kohonen map learning algorithm [2]. When learning stabilized, the resulting is the *latent semantics manifold*.

In order to form a query, we begin with the terms entered by the user, which we call the *inquiry*, composed of a set of keywords, a sentence, even a whole paragraph. We process it by stop words removal and stemming. The result is a series of stems *(keywords)* $Y = \{t_{k_1}, \ldots, t_{k_q}\}$. For the sake of generality, we assume that the user associated weights $\{u_{k_1}, \ldots, u_{k_q}\}$ to these terms. The inquiry can thus be represented as a point q in the word space. The inquiry modifies the context by subjecting it to a sort of partial learning. Let $[*]$ be the neuron in the map closest to the inquiry point q. The map is updated, through a learning iteration, in such a way that the neuron $[*]$ gets closer to the point q by a factor ϕ, with $0 < \phi \leq 1$. This is the *target context* $[\mu\nu]'$ of the query. The *target semantics* for our query is given by the difference between the target context and the original one: $[\tilde{\mu\nu}] = [\mu\nu] - [\mu\nu]'$ The values $[\tilde{\mu\nu}]$ in a neighborhood of $[\tilde{*}]$ constitute our complete query expression.

3 Context evolution

The test procedure of the following experiments is incremental. Each user starts with an empty context, a target context (either computing or neurophysiology), and a list of queries. Each "run" centers on a single query. The user is given the query word and uses it to retrieve documents using our system based on the *google*® search engine [4]. Out of the reults returned by the search engine, the user chooses a number of them that she considers relevant. These documents are used as an input to the procedure described in the previous section to build a *level 1* context. The same query is now repeated with the level 1 context, its precision is noted, and the results that the user downloads are added to the context to create a *level 2 context*. The process is then repeated: the same

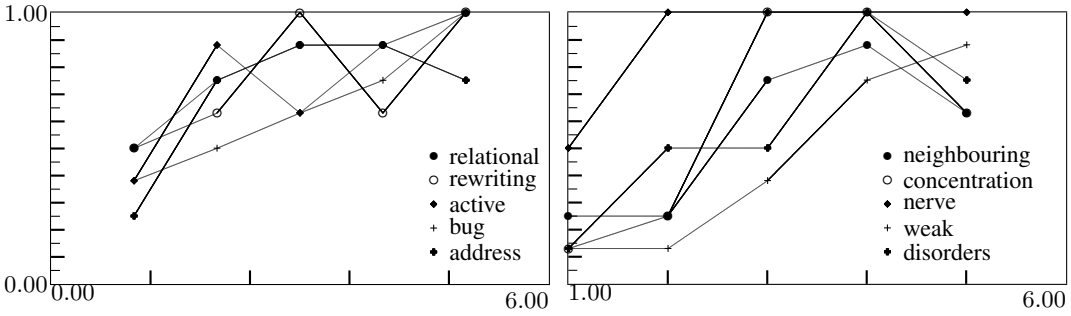

Fig. 2. Precision of the results for the computing context (left) and the neurophysiology context (right) as a function of the level of context.

query is sent using the level 2 context, and its precision is noted. The results of this first test are shown in figure **??**. The number n in abscissa is the number of results that we consider when measuring the rpecision, and the ordinate is the precision of the first n results. Note that the inclusion of level 1 context results in a significant improvement, while the inclusion of level 2 context leaves the results statistically unchanged for $n > 3$ and it appears to decrease the precision of the first two results. From an analysis of the documents retrieved, it seems that, with the second iteration, the context is somehow expanded. That is, at least with the users considered, the level 2 context appears to be less specific than the level 1, leading to the loss of precision.

In order to analyze the phenomenon further, we selected a few queries and repeated the procedure up the *level 4* context. The results are shown in figure 2 We can notice that the presence of "bumps" in the context is quite frequent, although the general trend, as expected, it towards an increase in precision with the context level. Here, too, we observe a phenomenon that we already pointed out in [4]: certain queries, mainly in the neurophysology context, are very specific and carry with them enough context information that adding context results only in marginal improvement. The query "nerve" is an example of this phenomenon. The precision seems to plateau after three or four iteration, indicating that at that time the context is quite formed and sufficient to specify the semantics of the desired query.

References

1. 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.
2. T. Kohonen. *Self-organizing maps*. Heidelberg, Berlin, New York:Springer-Verlag, 2001.
3. Simone Santini. Ontology: use and abuse. In *Proceedings of AMR 2007: international workshop on adaptive multimedia retrieval*. Heidelberg:Springer-Verlag, 2007.
4. Simone Santini and Alexandra Dumitrescu. Context as a non-ontological determinant of semantics. In *Proceedings of the third international conference on semantics and digital media technologies*, volume LNCS 5392, pages 121–137. Heidelberg:Springer-Verlag, 2008.