SEMANTIC CORRELATION IN CONTEXT:
APPLICATION IN DOCUMENT COMPARISON

GUY A. BOY
Artificial Intelligence Research Branch
NASA Ames Research Center
Mail Stop 244-17
Moffett Field, CA 94035

ABSTRACT

This paper outlines an approach to semantic correlation in context within the electronic documentation domain. It addresses the problem of quantifying resemblance or similarity between knowledge-based systems (KBSs) such as active documents. An important aspect of correlation is that it depends on context. Correlation between two KBSs can be defined in terms of common knowledge between them (semantic correlation) in context. Context can be physical, verbal, or historical. An example of semantic correlation in context is developed. In this example, taken from the Computer Integrated Documentation (CID) project, active documents are considered as KBSs able to advise the user of what to do next in the current situation. Potential research issues are discussed.

RÉSUMÉ

Cet article esquisse une approche à la corrélation sémantique en contexte dans le domaine de la documentation électronique. Il pose le problème de la quantification de la ressemblance ou similarité entre systèmes à base de connaissances (SBCs) tels que des documents actifs. Un aspect important de la corrélation est qu'elle dépend du contexte. La corrélation entre deux SBCs peut être définie en termes de connaissance commune (corrélation sémantique) en contexte. Le contexte peut être physique, verbal ou historique. Un exemple de corrélation sémantique en contexte est développé. Dans cet exemple, provenant du projet Documentation Electronique Intégrée, des documents actifs sont considérés comme des SBCs capables de conseiller l'utilisateur sur ses possibilités d'action dans la situation courante. Des orientations potentielles de recherche sont discutées.

1. INTRODUCTION
Correlation is a common topic in statistics, decision theory, and signal processing. Artificial intelligence (AI) has used and extended several theories and methods from these fields. For instance, Markov and Bayesian networks have been used for representing probabilistic knowledge to express informational dependency. Shafer has said that "probability is not really about numbers; it is about the structure of knowledge" (Pearl, 1988). The goal of this paper is to present some original aspects of correlation in the intelligent interaction domain. In particular, the problem of resemblance or similarity between knowledge-based systems (KBSs) is developed. In this paper, KBSs use a knowledge representation, called knowledge blocks, to store and use indexing knowledge (Boy, 1991a). Their own knowledge is learned from past experience. Correlation between two KBSs can be defined in terms of common knowledge between them (semantic correlation) in context. Intelligent interaction between KBSs and users remains a central issue.

This paper is built around two concepts of correlation: the correlation between data (assuming that there is no available domain theory), the correlation between models (assuming that there is at least a preliminary domain theory available). Data are the first level of entities on which a correlation can be made. For a long time, normative measures have been developed for data correlation (Snedecor, 1946). Models are generally conceptual structures that have been built to map physical or mental objects, as well as their relationships. To correlate two models, one has to observe outputs of both systems for the same inputs, and devise the best method to measure the degree of correlation between them. Correlation of models implies correlation of data. Models that we are considering are KBSs. In this research, KBSs are considered as cognitive agents that are able to interact with the environment.

This paper is structured as follows. In section 2, we introduce the issue of context in correlation. Section 3 presents an example of semantic correlation in context from the Computer Integrated Documentation (CID) project currently under development at NASA. Within CID, active documents are considered as a special kind of KBS. In the construction of a very large documentation, for instance, sometimes two similar documents can be generated by independent teams. The problem is then to check if different documents are describing the same things. Correlation between document descriptions has been investigated by several authors (Jardine and van Rijsbergen, 1971; van Rijsbergen, 1979; Griffiths et al., 1984; Voorhees, 1985; Raghavan and Wong, 1986; Salton, 1989). Feedback has been analyzed in information retrieval (Harper and van Rijsbergen, 1978). However, no exhaustive research has been carried out on correlation of documents in context. Section 4 discusses various issues on semantic correlation within the intelligent interaction framework, and proposes some of the potential research issues.

2. SEMANTIC CORRELATION AND CONTEXT

2.1. A FIRST STEP TOWARDS SEMANTIC CORRELATION

The Larousse French dictionary gives the following definitions for correlation:

(1) Reciprocal relation between two things;
(2) Relationship between two terms in which one is logically associated with the other.

2.1.1. Reciprocal relation between two things
In the first definition, the word "things" can be instantiated by: "data sets", or "models". Reciprocal relation between two data sets is generally handle by statistical methods. Data are usually represented by numbers. A reciprocal relation between two models is generally a combination of reciprocal relations between attributes of both models. These attributes describe the model, and may be models themselves.

If models are KBSs, they can be considered as cognitive agents. A common question in psychology is how to measure the cognitive properties of an agent? Physical things are directly perceived through their properties or attributes. Cognitive properties (or attributes) are more difficult to perceive and then to measure. Cognitive properties are measured by inference using a model and observable properties (behavior). For instance, we can infer workload by measuring someone's heart rate, or by recording subjective ratings from the user. A reciprocal relation between two KBSs can be defined in terms of a semantic distance of their cognitive properties. For instance, in the CID domain, active documents (referents) have cognitive properties such as contextual links that are relations between descriptors and referents in context (a formal definition is given in section 3.1.2 of this paper). These descriptors are either directly extracted from text (normative automatic indexing) or subjectively given by users.

2.1.2. Relationship between two terms in which one is logically associated with the other

In this second definition, the word "terms" can be instantiated by "concepts" or "descriptors". This leads to the construction of hierarchies of concepts. Many authors have already developed theories and methods in the field of classification and clustering (Duda & Hart, 1973). Machine learning techniques have been developed such as learning from examples (Quinlan, 1986) or concept clustering (Fisher, 1987). Recently, we have introduced the concept of context in such processes (Boy, 1991b). To illustrate the relationship between two terms, in which one calls logically the other, let us take the term "office". For instance, a typical thesaurus would give the following statement:

The term "office" is logically associated with the term phrases "special duty", "ceremonial observance", and "business place".

We know that "special duty" and "business place" are three disjunctive term phrases. Thus, the previous statement can be modified as follows:

The term "office" is logically associated with the term phrase "special duty" in the context of position of authority, "ceremonial observance" in a religious or social context, and "business place" in the context of location.

But to say that, we have to know the current context of discourse. Thus, in this case, context is specified by domain knowledge. Since correlation is generally dependent on context, the next section develops the concept of context.

2.2. DEFINING CONTEXT

The concept of context is not easy to define. Context can be related to the persistence of a situation. A generic situation is constructed from incremental integration (and generalization) of
various specific situations. For instance, words are defined by incrementally using them in various specific situations. Even if you do not know what the word *wishbone* means, the following sentences will contribute to its definition:

He used the best *wishbone* available in the race... This very short *wishbone* allowed him to turn shorter than the others at the marker buoy... Another reason is that such a *wishbone* is attached very high on the mast... In fact, all windsurf boards having such a *wishbone* should be used by real professionals...

After the first sentence, we learn that a *wishbone* can be used in a race. Even if we do not know what kind of race, this eliminates the possibility of an animal bone. In the second sentence, we learn that the *wishbone* is used on the water because of the marker buoy. After hearing the third sentence, we know that the *wishbone* is on a boat or some floating device. Finally, in the last sentence, we learn that a *wishbone* is part of a windsurf board (we assume that we know what a windsurf board is, otherwise we will continue to identify this new word in following conversations).

Thus, by adding situations (context) in which a word is used, this word can be incrementally defined. Such a context is called verbal context. This means that a word defined in a verbal context is defined by the other words around it. In information retrieval, the keyword in (verbal) context (KWIC) has been used as visual correlation during the selection (Luhn, 1959). This technique has been recently automated by Zimmerman (1988) and is used in CID.

There is another form of context that is physical and social. If you are presented with a windsurf board for instance, its owner may explain to you how to use it. In the conversation, he may say, "If you take the *wishbone* that way, the wind will push the sail in that direction". You suddenly discover that this piece of light metal supporting the sail laterally is called the wishbone. You learn this in a physical context. It may even happen that you will remember this word in this context forever, i.e., each time you think about your friend demonstrating his windsurf board, you will remember the anecdote of the wishbone, and conversely when you will think about wishbone. This is to say that context is very important for remembering things. In particular, in information retrieval, we claim that if context is used appropriately, it will enhance the search in memory. Hayakawa (1990) says that:

> The "definitions" given by little children in school show clearly how they associate words with situations. They almost always define in terms of physical and social contexts:
> "Punishment is when you have been bad and you have to sit on the stairs for a time out."
> "Newspapers are what the paperboy brings."

In more formal terms, relations between of objects can define context. An object can be a word or phrase, a picture, a movie sequence, a physical object (such as a table), a person, etc. These attribute/value pairs may be descriptions of verbal context (what is around the object and defines it in some situations), physical and social context (what one can point out in the environment of the object to concretize its definition), or historical context (what happened before and is causally related to the object being described). For instance the context of an information retrieval search could be described by a user profile (user model), e.g., the type of user, the type of task he/she is currently doing, the time, etc. The main problem with formal definition of context is that it can lead to a very large set of attribute/value pairs. From a computational point of view, consequent pattern matching may become practically impossible. This leads to the idea that context is usually defined by default for the most frequently used attribute/value pairs.
structure. This structure may have some exceptions. For instance, when we use the word "chair", it denotes a physical device that is used to sit on. However, in the context of a conference, the "chair" is usually the person who is in charge of the conference.

As Hayakawa (1990) points out, words have *extensional* and *intentional* meaning. If a real object that you can see in the environment can describe a word, its meaning is extensional. For instance, if you point at a cow in a field, you denote the cow. If a word is described by other words to suggest its presence, then its meaning is intentional. For example, you may describe a cow by saying that it is an animal, it has horns, it lives on a farm, etc. Each object can be associate to a particular intentional description according to who is giving the description, when the description is given, etc. Again, context is a key issue when objects have to be described. It is a fallacy to claim that each object could be associated with a single word. Anyone who tried to retrieve documents in a library using keyword equations knows this. It is almost impossible to retrieve the desired information because the description we give for the query seldom matches the descriptions that librarians have developed. However, if you talk with the librarian, he/she may help you better by acquiring more context from you. In the best case, if the librarian is a good friend and knows your work and needs, then he/she will be very helpful. If the librarian does not know you, he/she can capture physical context by simply looking at you. He/she may consider facts such as: you are young; you wear a lab coat; you knew where to ask in the first place; etc. He/she will also capture context from what you say (verbal context). Hayakawa says that "an examination of the verbal context of an utterance, as well as an examination of the utterance itself, directs us to its intentional meanings; an examination of the physical context directs us to the extensional meanings."

Another type of context should be mentioned: *historical context*. Context includes a temporal aspect, i.e., context summarizes a time period. Furthermore, persistence of some events reinforces context. For instance, a user told the librarian that he/she is looking for some information on geometry, he/she is currently involved in a computer class, he/she has a problem of drawing curves using a set of points, he/she needs to obtain a continuous drawing on the screen, and later he/she finally requested a reference on splines. The librarian usually integrates historical context, and, in this specific example, will not be confused between mathematical spline functions (what the requester needs) and the physical splines that draughtsmen were using in "ancient" times. In this example, the words "computer" and "screen" help to decide that the requested "splines" will be used in the last quarter of the 20th century.
3. AN EXAMPLE OF SEMANTIC CORRELATION IN CONTEXT

One of the most promising areas of research in knowledge acquisition is knowledge elicitation from text. Within the NASA Computer Integrated Documentation (CID) project, we have developed methods to elicit knowledge from text. CID is a system that enables integration of various technical documents in a hypertext framework (Boy, 1991c). It includes an intelligent browsing system that incorporates semantic indexing in context. Context can be physical (e.g., type of user, type of work he/she is doing, etc.), verbal (e.g., logical equations of keywords), or historical (e.g., time period attached to the request).

3.1. INDEXING AND INFORMATION RETRIEVAL IN CONTEXT

Indexing can be seen as a knowledge acquisition process from text and graphics. Information retrieval is the process of correlating descriptive queries with indexed documents. Both indexing and information retrieval depend on context. For instance, librarians and users may not use the same context when they index or try to retrieve documents.

Let us assume that each document (a book, a chapter, a paragraph or even a word) has its own knowledge base about its content. The main problem is to elicit this knowledge from them. Documents associated with their knowledge are called active documents. Active documents are considered as KBSs in the sense that they can "provide the right information at the right time in the right format according to the current context" or "direct the user towards other documents that are more correlated to his/her needs". Correlation between documents is not only a way to know if documents are related, but also a way to validate their elicited knowledge. This is the usual way that people validate their knowledge, by comparing it to the knowledge of others.

A thesaurus is the knowledge attached to whole documentation. It can be represented as a semantic network of documentation descriptors. When we want to validate a thesaurus, we usually correlate its knowledge with experts' knowledge. Such a validation depends on situations and points of views (i.e., context). In order to more formally define this artificial society of active documents, we first need to define some important concepts: descriptors, referents, and contextual links.

3.1.1. Descriptors and Referents

Indexing is the process of building descriptors \{d\} (descriptions) from referents \{r\} and linking them together in a given context (r->d). We call a descriptor any piece of text (word, sentence, or paragraph) or image (marked area or label on part of an image) that describes objectively or subjectively any other piece of documentation. Descriptors can be single term or multiple terms. We call a referent any piece of documentation (word, line of text, paragraph, picture, moving video or animated sequence, program, volume, library) that is described by at least one descriptor. A referent is always characterized by an objective identifier, but they also can have subjective identifiers. Let r be a referent and \{d_1, d_2, ..., d_n\} a set of descriptors that independently describe r. There must be at least one \(d_i\) that uniquely describes r. For instance, the address of this referent is such a descriptor. Generally, the other descriptors (other than \(d_i\)) are added because they are more convenient for users, even if they do not uniquely describe r. For instance, in the first page of this paper there is a referent called "Abstract". It can be uniquely described by "\(d_i=Abstract\)" and other descriptors (partial descriptions) such as "semantic correlation" and "context". Descriptors are not necessarily included explicitly in the content (text or graphics) of a referent. For instance, "Document comparison" is an implicit partial description
of the abstract of this paper. A referent can be any part of the documentation including the table of contents and the index. Tables of contents or indexes are usually found more convenient because of their well known structure (hierarchical and alphabetical).

*Information retrieval* is the process of retrieving referents from available descriptors in a given context (d->r). Information retrieval is a process of *abduction* using indexing knowledge, in that knowing r->d and d, r becomes a hypothesis, i.e., in practice r becomes a possible referent that the user may consider. We say that indexing defines a set of semantic relations that are used in information retrieval. Figure 1 shows an example of semantic relations between referents via descriptors.

### 3.1.2. Contextual Links

Generally, knowledge about indexing and information retrieval is not explicit and constitutes human expertise. If the goal is to help the user browsing and retrieving appropriate information, such knowledge should be made explicit and implemented in an intelligent assistant system, as described in (Boy, 1991a), built on top of the hypertext. A major advantage of this architecture is that this metalevel is easily programmable and allows the inclusion of knowledge on the links of the hypertext.

We have shown that knowledge blocks adapt well for representing procedures used in document browsing (Boy, 1990). Thus, they are used to represent links between descriptors and referents. A knowledge block contains a goal, a set of conditions, and a set of actions to achieve the goals. Conditions are decomposed into triggering conditions, abnormal conditions, and contextual conditions. In CID, triggering conditions are represented by descriptors that the user selects. Contextual conditions may describe the external environment (e.g., type of user, type of task when doing the consultation of documentation, and time period), or internal history of the current documentation use (e.g., the last visited referents, the last selected descriptors). Actions are represented by referents where the user wants to go. A goal is reached when the user selects the success postcondition corresponding to the chosen referent. Similarly, an abnormal condition corresponds to a failure postcondition. Users browsing and formulating their descriptions of the content of referents contribute to the augmentation of the knowledge base by acknowledging success (or failure) postconditions after a successful (or unsuccessful) search.
In CID, knowledge blocks are used to represent *contextual links* between descriptors and referents. This aspect has been developed in Boy (1991c). In this view, documentation can be represented by two entities: referents and contextual links. As each referent \( r \) is described by a set of descriptors \( D(r) \), a contextual link between two referents \( r_1 \) and \( r_2 \) implicitly assume that \( r_1 \) and \( r_2 \) share at least one common descriptor. Then, we can define the set of semantically common descriptors. Figure 2 presents the integration of referents (text or graphics) with corresponding contextual links (knowledge). Each descriptor in a referent is a *semantic direction* towards a set of related referents. For instance, let us assume that the referent \( r_1 \) has two semantic directions (descriptors \( d_1 \) and \( d_2 \)) corresponding to the contextual links:

\[
\{ d_1; (r_2 \mid C_{12}), (r_3 \mid C_{13}) \} \text{ and } \\
\{ d_2; (r_4 \mid C_{14}), (r_5 \mid C_{15}) \},
\]

where \( \{ d_1; (r_2 \mid C_{12}), (r_3 \mid C_{13}) \} \) can be read as: the contextual link between the descriptor \( d_1 \) and the referent \( r_2 \), in the context \( C_{12} \), and the referent \( r_3 \), in the context \( C_{13} \).

If the current context is properly identified and the system knowledge is stabilized (i.e., sufficient contextual links have been built), then contextual links help the user to decide the action which is currently most appropriate. For instance, if the user is facing the referent \( r_1 \), then the system proposes two possible semantic directions, \( d_1 \) and \( d_2 \), if the user selects \( d_1 \) the system proposes to go to \( r_2 \) or \( r_3 \) according to the current context. In this sense, a referent and its associated knowledge base is called an *active document* (i.e., a *documentation agent*) that has the capability of helping the user select his next move. The first major issue in this approach is the acquisition of general contextual conditions (Boy, 1990). A second important issue is to help the user finding his way in the documentation quickly. In the next subsection, we will introduce the main problem of semantic correlation between referents as a navigation aid.
3.2. SEMANTIC CORRELATION BETWEEN ACTIVE DOCUMENTS

Indexing is improved if users' needs are taken into account in index generation and maintenance, i.e., if each active document has an appropriate user model. Generally, users refine (and often define) their needs by trial and error evaluation of information retrieval results. Such behavior necessitates an interactive environment that allows incremental acquisition of indexing knowledge.

Indexing is a decision making process that must be directly accessible to the user. This decision process involves building semantic relations, i.e., relations between descriptors, between referents, and between descriptors and referents. The corresponding knowledge representation used to handle this decision process must allow repair in case of failure, i.e., when the user did not make the right decision. It should allow incremental transformation.

3.2.1. Incremental reinforcement in context

Contextual links can be seen as user-generated traces in the documentation. According to the ACT theory (Anderson 1976), a trace once formed is not lost, but its strength may decay. By analogy with human, long-term retention show gradual but continuous forgetting. Based on data summarized by Wickelgren (1976), trace strength is a power function of time. Using these psychological results, we have chosen a simple heuristic [formula (1) below] to record a minimal set of parameters during the use of the system. We have noticed that the relevance of a contextual link \{d_i, (r)\} between a descriptor d_i and a referent r depends on user's feedback (success or failure) on r, the frequency of feedbacks on r, and the importance that the user...
assigns to this referent \( r \). The semantic relevance \( \text{Rel}(r \mid d_i, C) \) of a referent \( r \) with respect to a descriptor \( d_i \) in context \( C \) has been formally defined (Boy, 1991b) as:

\[
\text{Rel}(r \mid d_i, C) = \sum_{s \in S_{r \mid d_i, C}} I_s (t-t_s)^{-\beta_s} - \sum_{f \in F_{r \mid d_i, C}} I_f (t-t_f)^{-\beta_f}
\] (1)

where \( t \) is the current time, \( t_s (t_f) \) is the time when the user selected the success \( s \) (failure \( f \)), \( S_{r \mid d_i, C} (F_{r \mid d_i, C}) \) is the set of successes (failures) through time \( t \), the exponent \( \beta_s (\beta_f) \) has a value on the interval 0 to 1, and expresses the decay from success (failure), and \( I_s (I_f) \) is the importance assigned to \( r \) by the user with respect to \( d_i \) in context \( C \). In practice, \( I_s, I_f, \tau_s, \) and \( \tau_f \) may be set up as constants for all users' feedbacks. This heuristic simulates the fact that contextual links can be forgotten if they are not reinforced after a long period of time, except if a large importance coefficient \( (I_s \text{ or } I_f) \) has been assigned to them.

3.2.2. An example of semantic correlation

In the Space Station documentation application, documents (referents) are generated by people who are not necessarily in contact with one another. These documents are theoretically connected through a top-down hierarchical structure (table of contents). However, there are no transverse links anticipated in the current documentation system. This is a classical problem in the construction of a very large documentation; sometimes two similar referents can be generated by independent teams. The problem is then to check if different referents are describing the same things. If this is the case, can they be merged into a single referent? In order to solve this problem, it is necessary to generate well-defined descriptors that can be shared by other documentation writers. If these descriptors are well defined, we can expect to check the degree of interdependence between referents by measuring their semantic correlation.

A semantic correlation between two referents should express the resemblance between the content of each referent. It can be constructed taking into account the descriptors attached to each referent. These descriptors characterize dimensions along which each referent can be located in the descriptor space. If descriptors express the semantics of the referent space, then the descriptor space can be called semantic space by extension. The main problem is that these dimensions are not independent. As already explained, the descriptor space represents relations between descriptors that express inheritance or property relations. Furthermore, it seems reasonable to let people index their referents without constraining them with the problem of descriptor dependence. The problem is then to design a representation that makes descriptor dependencies explicit. The basic idea is to build a semantic network where nodes are descriptors, and links are hierarchical (inheritance) or property links. Taking this approach, descriptors are organized into descriptor dependency clusters. These clusters may be loosely connected or actually independent.

We consider that the following parameters affect semantic correlation in context between two referents \( r_1 \) and \( r_2 \):

- the number of descriptors \( n_{db}(r_1, r_2) \) shared between \( r_1 \) and \( r_2 \),
- the semantic relevance of a referent with respect to a descriptor of the shared set in a given context (derived from user feedback),
- the hierarchical level of the shared descriptors in each referent,
- the current context.
Let us assume that we want to compare two referents $r_1$ and $r_2$. Salton (1989) defines the similarity between two referents $r_1$ and $r_2$ by a function of the number of descriptors shared by both referents. The referents $r_1$ and $r_2$ have two descriptor lists $D(r_1)$ and $D(r_2)$. The current context is taken into account by restricting $D(r_j)$ [with $j=1$ or 2] to the list of descriptors that have been "sufficiently" reinforced [$\text{Rel}(r_j \mid d_i, C) > \text{Threshold}$] in the current context. The intersection between $D(r_1)$ and $D(r_2)$ is then computed using the following rules:

If $d \in D(r_1)$ and $\exists d' \in D(r_2), \exists n$, such that $d' = \text{childOf}^{(n)}(d)$
then $d \in D(r_1) \cap D(r_2)$ \hspace{1cm} (3)

If $d \in D(r_1)$ and $\exists d' \in D(r_2), \exists n$, such that $d = \text{childOf}^{(n)}(d')$
then $d' \in D(r_1) \cap D(r_2)$ \hspace{1cm} (4)

where the function $d' = \text{childOf}^{(n)}(d)$ means that $d'$ is a child of generation $n$ of $d$. For instance, $\text{childOf}^{(0)}$ would be identity function, $\text{childOf}^{(1)}$ would be the direct child function (first generation), $\text{childOf}^{(2)}$ would be the grand child function (second generation), etc. In other words, we keep the most general descriptors in the intersection.

If the set $D(r_1) \leftrightarrow D(r_2)$ is not empty, then it can be ordered with respect to their corresponding semantic relevances. We obtain the following table:

<table>
<thead>
<tr>
<th>$D(r_1) \leftrightarrow D(r_2)$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>rank$_{11}$</td>
<td>rank$_{12}$</td>
<td>rank$_{13}$</td>
<td>...</td>
</tr>
<tr>
<td>$r_2$</td>
<td>rank$_{21}$</td>
<td>rank$_{22}$</td>
<td>rank$_{23}$</td>
<td>...</td>
</tr>
</tbody>
</table>

where rank$_{12}$ is the rank of descriptor $d_2$ in the subset $D(r_1) \leftrightarrow D(r_2)$ of the descriptors of $r_1$. Each rank$_{ij}$ is computed from the relative position of $d_i$ in $r_j$ with respect to the relevance $\text{Rel}(r_j \mid d_i, C)$. For instance, the above table shows the following order:

$\text{Rel}(r_1 \mid d_1, C) \leq \text{Rel}(r_1 \mid d_2, C) \leq \text{Rel}(r_1 \mid d_3, C) \leq ...$

$\text{Rel}(r_2 \mid d_1, C) \leq \text{Rel}(r_2 \mid d_2, C) \leq \text{Rel}(r_2 \mid d_3, C) \leq ...$

The Spearman rank correlation coefficient applied to this problem gives the following formula (Snedecor, 1946) for semantic correlation in context:

$$
\rho_{\text{Sem}}(r_1, r_2 \mid C) = 1 - \frac{6 \sum_{d \in D(r_1) \cap D(r_2)} [\text{rank}_{1,i} - \text{rank}_{2,i}]^2}{n_{ds}(r_1, r_2)^3 - n_{ds}(r_1, r_2)}
$$

(5)
where \( n_{ds}(r_1, r_2) \) is the number of descriptors shared between \( r_1 \) and \( r_2 \), the following relation: \( 1 \leq \text{rank}_{ji} \leq n_{ds}(r_1, r_2) \) holds for \( j=1 \) or \( 2 \), and \( \rho_{\text{Sem}}(r_1, r_2 | C) \) can range from -1 (complete discordance) to +1 (complete concordance).

This semantic similarity measure can be used as a navigation aid. Two generic cases can be described as derivations of ordered lists of referents semantically correlated to a given referent, or a complex query from the user.

1. From a referent \( r_1 \), \( \rho_{\text{Sem}}(r_1, r_2 | C) \) may be computed for all the referents \( r_2 \) that share descriptors with \( r_1 \).

2. A complex query involving the conjunction of several descriptors can be represented by a set \( D(Q) \). An analogous formula \( \rho_{\text{Sem}}(Q, r | C) \) can be derived for any referent \( r \) that shares at least one descriptor with \( Q \).

In both cases, the corresponding ordered list of referents can be presented to the user for validation purposes, or as a suggestion of what to do next.

In the acquisition of contextual links, it can be used to fine tune referent descriptions. For instance, if two referents \( r_1 \) and \( r_2 \) are highly correlated but the user does not agree with this result, then referent descriptions \( D(r_1) \) and/or \( D(r_2) \) have to be revised. In future consultations, this revision will apply to this particular user, except if the revision has been done for a class of users.

4. DISCUSSION AND POTENTIAL RESEARCH ISSUES

There is still much to be done to define semantic correlation in context. This paper outlines an approach to this problem. We have seen a way of defining correlation between KBSs as a measure of common cognitive properties they shared in context. These cognitive properties are reinforced in context. An example has been presented in the CID domain.

Semantic correlation in context is a crucial concept for intelligent interaction between agents. In particular, analysis, design, and evaluation of human-machine intelligent interfaces can be greatly improved by definition of cognitive measures between human and machine agents. Examples are provided in (Boy & Tessier, 1985). Such measures are difficult to define because they involve psychological properties. Such properties are difficult to elicit and perceive. They are usually inferred from a model and observable behaviors.

As an agent develops expertise, he/she develops situational knowledge, i.e., the ability to recognize things easily. In other words, correlation between an agent's knowledge and the situation depends on the agent's level of expertise or stage of learning. Even if analytical knowledge and experience contribute to improve situational knowledge, there is a long way to go before a machine will be able to generate such knowledge automatically. This is the reason why, in the CID system, we have decided to have human agents provide such knowledge on-line when they are interacting with active documents. The awareness of correlation between documents increases the perception by human agents of internal interaction between active documents. An interesting research issue is to verify if user models imbedded in active documents increase interaction with users.

Object-oriented programming is now fairly widely used in computer science, and developers started to think in terms of objects and agents. It will not be long before designers will think in
terms of societies of artificial agents for designing new architectures. Such cognitive primitives (Rappaport, 1987) will be integrated as we integrate lines of code today. Again, when the designer suggests an intention, cognitive measures will be necessary to retrieve appropriate artificial agents (in libraries). Such correlations between the intention of an agent and capabilities of other agents remain to be defined.

Correlation does not depend only on correlation formulas and models. Knowledge is necessary to get successful and accurate correlations, and this knowledge may come from the domain that is analyzed or from the way correlation is processed. In other words, the human correlator is a very important factor in the success of the correlation process. For instance, an expert in circadian rhythms, who analyzes physiological data recorded during flights and uses statistical packages to extract correlations between various physiological parameters, must have a background in statistics, physiology of human rhythms, and aviation. The more extensive this background, the more frequently the correlation process will lead to interesting results. In this particular case, people try to infer causal relations between schedules and human performance and workload. Interpreting correlations to induce causality is a very delicate process that requires expert knowledge at least in the three domains mentioned above. If such expert knowledge has to be acquired to implement an expert system that would be used as an intelligent assistant for a less expert correlator in the correlation process, the cognitive behavior of such a system must be correlated to the expert himself/herself. This process of correlation is generally called validation in the AI community. Validation of expert systems is still a topic of research. The concept of semantic correlation between agents can be also applied to the problem of expert system validation, that is, the correlation between experts expectations and the performance of the system, and correlation between the evaluations of different experts.

In conclusion, as we can expect that smarter computers and machines will become part of daily life in the next century, research on interaction with such machines should be intensified. These tools should be able to observe human behavior and infer their intentions. Work has been already started in this direction in various pilot associate programs. Semantic correlation in context may help to assess the degree of concurrence or cooperation between agents working together, when one agent is a machine and the other is a human. Development of hypermedia technology combined with the development of natural language processing opens the door to intelligent interaction. This cannot occur without continuing research on context to better define it, represent it, and acquire it. These are questions that will find better answers in the future.

ACKNOWLEDGEMENTS

I owe thanks to Robin Hanson who helped me a great deal in focusing this paper on the context-sensitive part of semantic correlation. Philippa Gander, Nathalie Mathé, Yaron Gold, Peter Cheeseman, and Pat Langley provided astute advice towards improving the quality of this paper.

REFERENCES


