Probabilistic Models for Semantic Representation

Francesco Colace, Massimo De Santo and Paolo Napoletano

University of Salerno, DIIIE
84084 Fisciano, Salerno, Italy

Abstract. In this work we present the main ideas behind *in Search of Semantics* project which aims to provide tools and methods for revealing semantics of human linguistic action.
Different part of semantics can be conveyed by a document or any kind of linguistic action: the first one mostly related to the structure of words and concepts relations (*light* semantics) and the second one related to relations between concepts, perceptions and actions *deep semantics*. As a consequence we argue that semantic representation can emerge through the interaction of both.
This research project aims at investigating how those different parts of semantics and their mutual interaction, can be modeled through probabilistic models of language and through probabilistic models of human behaviors.
Finally a real environment, a web search engine, is presented and discussed in order to show how some part of this project, *light* semantics, has been addressed.

1 Introduction

Semantic knowledge can be thought of as knowledge about relations among several types of elements: words, concepts, percepts and actions [1]. Formalization efforts to capture such aspects have been splitted in two different approaches. A first approach has focused more on the structure of associative relations words-words in natural language use and relations words-concepts, which we may define as *light semantics*. A second one has emphasized abstract conceptual structure, focusing on relations among concepts and relations between concepts and percepts or actions, which we may call *deep semantics*. However, these different aspects are not necessarily independent and can interplay to influence behavior in different ways. Thus, we will assume here that semantic representation can emerge through the interaction of both *light* and *deep semantics*.

The project aims at investigating how *light* and *deep semantics* -and their mutual interaction - can be modeled through probabilistic models of language and through probabilistic models of human behaviors (e.g., while reading and navigating Web pages), respectively, in the common framework of most recent techniques from machine learning, statistics, information retrieval, and computational linguistics. Such a model could handle each part of semantics and the cooperation among them in order to reveal intensions, meanings, or more properly semantics in a broader sense.

The paper is organized as follows. In Section 2 we introduce basic notion about semantic representation. A viable road to semantics (the core of *in Search of Semantics* - *iSoS* framework) is presented in Section 3, in Section 4 methods for handling light

Colace F., De Santo M. and Napoletano P. (2009).
Probabilistic Models for Semantic Representation.
In *Proceedings of the 1st International Workshop on Ontology for e-Technologies OET 2009*, pages 13-22
DOI: 10.5220/0002222100130022

semantics are presented and a real environments is introduced and discussed in Section 5, where some experiments are presented. Finally, in Section 6 we discuss conclusions and future works.

2 In Search of Semantics

The Semantic Web and Knowledge Engineering communities are both confronted with the endeavor to design and build ontologies by means of different tools and languages, which in turn raises an “ontology management problem” related to the peculiar tasks of representing, maintaining, merging, mapping, versioning and translating [2].

These mentioned above are well known concerns animating the debate in the ontology field. However, we argue that the utilization of different tools and languages is mainly due to a personal view of the problem of knowledge representation, which in turn raises a not uniform perspective.

Most important each ontology scientist may rely, deliberately or implicitly, on a different definition of the role of ontology as mean for semantics representation [3]. Therefore we argue that a special effort should be devoted to better explain and clarify the theory of semantic knowledge and how we should correctly model the latter for being properly represented and used on a machine.

A simple process to convey meaning through language can be summarized as follows:

meaning → encode → language → decode → *meaning’*,

where, since encoding/decoding processes are noisy, *meaning’* is the estimation of the original *meaning*. In order to understand why those processes are noisy we assume that a communication act through language is in the form of writing/reading a book. Here, the origin of the communicative act is a meaning that resides wholly with the author, and that the author wants to express in a permanent text. This meaning is a-historical, immutable, and pre-linguistic and is encoded on the left-hand side of the process; it must be wholly dependent on an act of the author, without the possibility of participation of the reader in an exchange that creates, rather than simply register, meaning. The author translates such creation into the shared code of language, then, by opening a communication, he sends it to the reader at the encoding stage. It is well known that, due to the accidental imperfections of human languages, such translation process may be imperfect, which in turn means that such a process is corrupted by “noise”. Once the translated meaning is delivered to reader, a process for decoding it starts. Such process (maybe also corrupted by some more noise) obtains a reasonable approximation of the original meaning as intended by the author. As a consequence meaning is never fully present in a sign, but it is scattered through the whole chain of signifiers: it is deferred, through the process that Derrida [4] indicates with the neologism differnce, a dynamic process that takes plane on the syntagmatic plane of the text [5].

In the light of this discussion we argue that, as pointed out by Steyvers and his colleagues [1], the semantic knowledge can be thought of as knowledge about relations among several types of elements, including *words*, *concepts*, and *percepts*. According to such definition the following relations must be taken into account:

1. *Concept – concept* relations. For example: knowledge that dogs are a kind of animal, that dogs have tails and can bark, or that animals have bodies and can move;
2. *Concept – action* relations: Knowledge about how to pet a dog or operate a toaster.
3. *Concept – percept* : Knowledge about what dogs look like, how a dog can be distinguished from a cat;
4. *Word – concept* relations: Knowledge that the word dog refers to the concept dog, the word animal refers to the concept animal, or the word toaster refers to the concept toaster;
5. *Word – word* relations: Knowledge that the word dog tends to be associated with or co-occur with words such as tail, bone.

Obviously these different aspects of semantic knowledge are not necessarily independent, rather those can influence behavior in different ways and seem to be best captured by different kinds of formal representations. As a consequence result, different approaches to modeling semantic knowledge tend to focus on different aspects of this knowledge, specifically we can distinguish two main approaches:

I The focus is on the structure of associative relations between words in natural language use and relations between words and concepts, along with the contextual dependence of these relations [6–8]. This approach is related to points 4 and 5, which can be defined as *light semantics*;

II The emphasis is on abstract conceptual structure, focusing on relations among concepts and relations between concepts and percepts or actions [9]. This approach is related to points 1, 2 and 3, which can be defined as *deep semantics*.

The key idea of this project is that indeed semantics representation is likely to emerge through the interaction of *light* and *deep semantics*. Thus, an an artificial system contending with semantics should necessary take into account both facets [2].

3 A Viable Road to Semantics

Once a computational model for each of the two components of semantics has been formulated, the very aim of this research project is to investigate the interaction between them and how such interaction can be modeled through probabilistic methods. In the following we will show how each relations, discussed in Section 2, can be modeled.

We argue that probabilistic inference is a natural way to address problems of reasoning under uncertainty, and uncertainty is plentiful when retrieving and processing linguistic stimuli. In this direction, it has been demonstrated that language possesses rich statistical structure that could be captured through probabilistic models of language based on recent techniques from machine learning, statistics, information retrieval, and computational linguistics [10].

Specifically, the description of both *Word – word* and *Word – concept* relations, namely *light semantics*, is based on an extension of the computational model, namely the topic model, introduced by Steyvers in [1] and [11]. Topic model is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifes a simple probabilistic procedure by which documents can be generated.

The deep semantics is traditionally represented in terms of systems of abstract proposition [9]. Models in this tradition have focused on explaining phenomena such as the development of conceptual hierarchies that support propositional knowledge, reaction time to verify conceptual propositions in normal adults, and the decay of propositional knowledge with aging or brain damage.

While *Concept – concept* relations could be modeled using the prototype theory that plays a central role in linguistics, as part of the mapping from phonological structure to semantics [12], most interesting for us, the *Concept – action* relations can be revealed using the theory of *emergent semantics* pointed out by Santini and Grosky [13, 14].

Building semantics by using perception (vision, etc.), that is the modeling of *Concept – percept* relations, is a problem that can be understood by considering the Marr's computational theory [15]. Here we will investigate the mechanism describing how the human make use of perception (in broad sense) for encoding knowledge representation. For instance could be interesting to investigating mechanisms of sensory-motor coordination in order to understand how such mechanisms can reveal deep semantics. In such perspective studies in the field of Computer Vision will be useful [16]. One of the approach that seems to be suitable for our purpose is that proposed by Pylyshyn [17] for situating vision in the world by differentiating three different ways in which an artificial agent might represent its world in order to carry out actions in real world.

In this scenario we are interested in designing a computational model for combining all these aspects of semantics and to account for user semantics.

4 Ontology Building in a Probabilistic Framework for Light Semantics Representation

4.1 Troubles with Ontology

Different approaches have been used for building ontology: manual, semiautomatic and automatic methods. Most of them are manual, however among the semiautomatic and automatic methods we can distinguish these based on Machine Learning techniques from these based on pure Artificial Intelligence theory [18]. Notwithstanding those considerations, the great majority of existing methods relies on a concept of ontology according to what is commonly acknowledged in computer science field, that is an ontology is a set of terms, a collection of relations over this set, and a collection of propositions (axioms) in some decidable logical system. In this direction, the Web Ontology Language (OWL) represents the most used language for authoring ontologies; it is, as declared by the W3C consortium: "a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is developed as a vocabulary extension of RDF (the Resource Description Framework)..."

By embracing the debate raised in [3], we rely on a on a different definition of the role of ontology as mean for semantics representation, more precisely in our opinion the ontology should abandon any velleity of defining meaning, or of dealing with semantics, and re-define itself as a purely syntactic discipline. Ontology should simply be an instrument to facilitate the interaction of a user with the data, keeping in mind that the user's situated, contextual presence is indispensible for the creation of meaning. For instance, it would be a good idea to partially formalize the syntactic part of the interaction process that goes into the creation of meaning.

By following this direction, it is our conviction that one of the major limitations of languages for representing ontologies - and in this respect OWL is no exception - stems from the static assignment of relations between concepts, e.g. "Man is a subclass of Human". On the one hand, ontology languages for the semantic web, such as OWL and RDF, are based on crisp logic and thus cannot handle incomplete, partial knowledge for any domain of interest. On the other hand, it has been shown how (see, for instance [19]) uncertainty exists in almost every aspects of ontology engineering, and probabilistic directed Graphical Models (GMs) such as Bayesian Nets (BN) can provide a suitable tool for coping with uncertainty. Yet, in our view, the main drawback of BNs as a representation tool, is in the reliance on class/subclass relationships subsumed under the directed links of their structure. We argue that an ontology is not just the product of deliberate reflection on what the world is like, but is the realization of semantic interconnections among concepts, where each of them could belong to different domains.

Indeed, since the seminal and outstanding work by Anderson on probabilistic foundations of memory and categorization, concepts/classes and relations among concepts arise in terms of their prediction capabilities with respect to a given contex [20]. Further, the availability of a category grants the individual the ability to recall patterns of behavior (stereotypes, [21]) as built on past interactions with objects in a given category. In these terms, an object is not simply a physical object but a view of an interaction.

Thus, even without entering the fierce dispute whether ontologies should or should not be shaped in terms of categories [22], it is clear that to endow ontologies with predictive capabilities together with properties of reconfigurability, what we name *ontology plasticity*, one should relax constraints on the GM structure and allow the use of cyclic graphs. A further advantage of an effort in this direction is the availability of a large number of conceptual and algorithmic tools that have been produced by the Machine Learning community in most recent years.

The main idea here is the introduction of a method for automatic construction of ontology based on the extension of the probabilistic topic model introduced in [1] and [23].

4.2 Ontology for Light Semantics

The description of both *Word – Word* and *Word – Concept* relations, related to the light part of semantics, is based on an extension of the computational model depicted above and discussed in [1] and [11]. Here we discuss how to model *Word – Word* relations, whereas the *Word – Concept* relations are modeled by using the concept-topic model proposed in [11]. We consider that, together with the *topics model*, what we call the *words model*, in order to performs well in predicting word association and the effects of semantic association and ambiguity on a variety of language-processing and memory tasks.

The original theory of Griffiths mainly asserts a semantic representation in which word meanings are represented in terms of a set of probabilistic topics z_i where the statistically independence among words w_i and the “bags of words” assumptions were made. The “bags of words” assumption claims that a document can be considered as a feature vector where each element in the vector indicates the presence (or absence) of a word and where information on the position of that word within the document is completely lost. Assume we will write $P(z)$ for the distribution over topics z in particular document and $P(w|z)$ for the probability distribution over word w given topic z. Each word w_i in a document (where the index refers to ith word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. We write $P(z_i = j)$ as the probability that the jth topic was sampled for the ith word token, that indicates which topics are important for a particular document. More, we write $P(w_i|z_i = j)$ as the probability of word w_i under topic j, that indicates which words are important for which topic. The model specifies the following distribution over words within a document:

$$P(w_i) = \sum_{k=1}^{T} P(w_i|z_i = k)P(z_i = k) \tag{1}$$

where T is the number of topics. In through the *topic model* we can build consistent relations between words measuring their degree of dependence, formally by computing joint probability between words:

$$P(w_i, w_j) = P(w_i|w_j)P(w_j) = \sum_{k=1}^{T} P(w_i|z_i = k)P(w_j|z_j = k) \tag{2}$$

Several statistical techniques can be used for unsupervised inferring procedure great collections of documents.

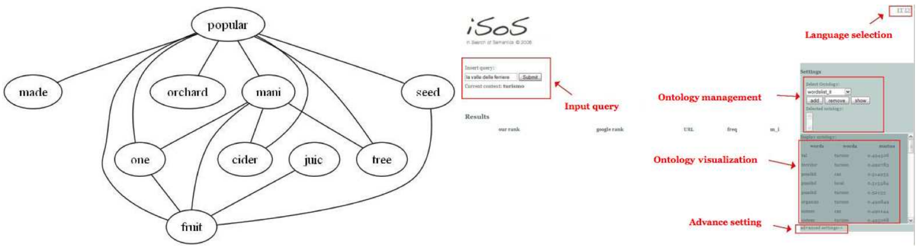
We will use a generative model introduced by Blei et al. [23] called latent Dirichlet allocation. In this model, the multinomial distribution representing the gist is drawn from a Dirichlet distribution, a standard probability distribution over multinomials. The results of LDA algorithm, obtained by running Gibbs sampling, are two matrix:

1. the words-topics matrix Φ: it contains the probability that word w is assigned to topic j;
2. the topics-documents matrix Θ: contains the probability that a topic j is assigned to some word token within a document.

By comparing joint probability with probability of each random variable we can establishes how much two variables (words) are statistically dependent, in facts the hardness of such statistical dependence increases as mutual information measure increases, namely:

$$\rho = \log|P(w_i, w_j) - P(w_j)| \tag{3}$$

where $\rho \in [0, -\infty]$. By selecting hard connections among existing all, for instance choosing a threshold for the mutual information measure, a GM for the words can be delivered, (cfr. Figure 1(a)). As a consequence, an ontology can be considered as set of pair of words each of them having its mutual informational value.



Insert query: la valle delle ferriere Submit

Current context: turismo

(a) (b)

Fig. 1. 1(a) Graphical model representing *apple* ontology, obtained from a set of documents about the topic *apple*. 1(b) in Search of Semantic web search engine's functionalities screenshot: querying, ontology visualization etc.

Building Ontology: A Case of Study of Light Semantic. Here we present a case of study of light semantics representation. Once topic is chosen, the words connections, namely *words model*, are learned from large text corpora, and consequently a Graphical model representing "apple" (as fruit) ontology is builded. The multidocument corpus, extracted from a web repository obtained by a craling stage for the query *apple*, the number of documents are 200. As a result, we show the GM representing light semantics relations for the "apple" topic: the ontology is showed in Figure 1(a).

5 Real Environment for Light Semantics

Probabilistic model of light semantics are embedded in a real web search engine developed at University of Salerno and reachable through the URL http://193.205.164.69/isos after a registration procedure. In Figure 1(b) is showed a screenshot for the *iSoS* web search engine and in following we describe its principal functionalities. We choose a web search engine as a laboratory where developing and testing methods for treating semantics, here we have documents created by human (Web pages) and the opportunity for tracking human behavior at same moment.

5.1 In Search of Semantics: Functionalities

As discussed above, *iSoS* is a web search engine with advanced functionalities. This engine is a web based application, entirely written in Java programming language and Java Server Page Language embedding some of the open source search engine Lucene [1] functionalities. As basic functionalities it performs sintax querying, see the left side of Figure 1(b), and it gives results as a list of web pages ordered by frequency of the term query.

The *iSoS* engine is composed of three parts: Web crawling, Indexing, Searching. Each web search engines work by storing information about web pages, which are retrieved by a Web crawler, a program which follows every link on the web. In order to

[1] http://lucene.apache.org/

better evaluate the performance of such web search engine, a small real environment is created. It performs a simplified crawling stage by submitting a query to a famous web search engine Google (www.google.com), and crawling the URL of the web pages contained in the list of results of Google. During the indexing stage each page is indexed by performing a traditional technique, the *tf-idf* schema. The searching stage is composed of 2 main parts. The first is a language parsing stage for the query, where stop words like “as”, “of ” and “in”, are removed and the second is a term searching stage in the *tf-idf* schema.

The *iSoS* web search engine provides advanced functionalities, namely the ontology builder tool, where the user can build ontology by exploiting the procedure discussed in Section 4. We can also manage ontology, Fig. 1(b), by adding each ontology to a list that we call the knowledge domain.

Users can decide to include the ontology in a simple query searching task. In this case, *iSoS* searches both the query terms set and the ontology pairs of terms. We argue that more relevant document can be retrieved by exploiting specific knowledge domain. The search engine provide some general ontology and, more interesting, allows each user to create own ontology by using the Ontology builder tool. This functionality is reserved to a registred user and is realized through a kind of user profiling technique.

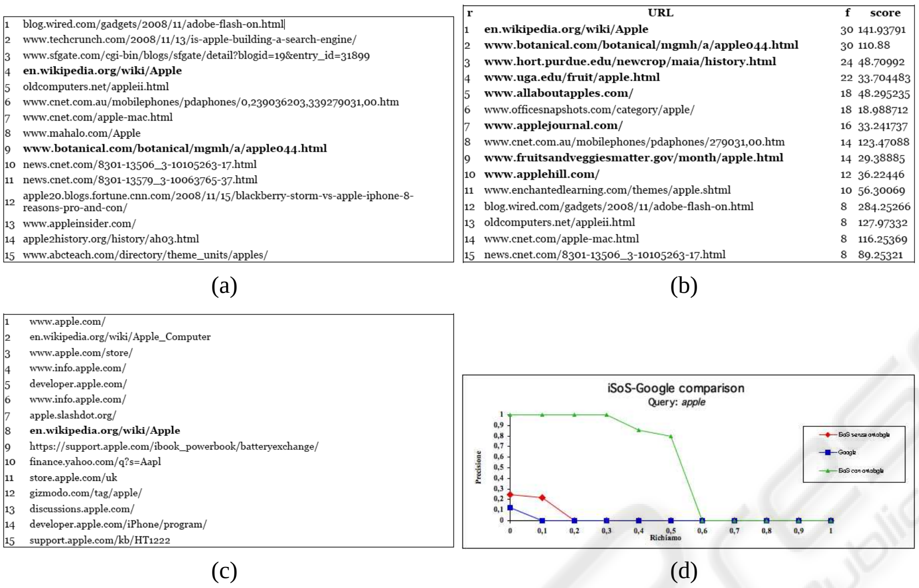
5.2 Experimental Results

In order to evaluate the performance of *iSoS*, we have indexed several web domain, *apple*, *bass* and *piano*, we have performed several term queries and finally we have compared with Google (a custom version of it). For each domain we have created a small web pages repository composed of 200 documents obtained by using the crawling procedure discussed above, and the ontology of apple is builded by using a small repository of documents collected by an expert of this topic. When an index is given we have submitted several query and computed the precision-recall graphs for both *iSoS* and Google. Due to problem of limit in space, here we discuss only results obtained for the domain *apple*.

***Apple* Domain.** This domain mainly contains documents about Apple inc., but we are interest in apple as fruit. To solve this ambiguation and to present most relevant results we make use of the ontology. In Fig. 2(a), 2(b), 2(c), are reported the first 15 results obtained with *iSoS* without ontology, *iSoS* with ontology and Google Inc. respectively, and finally in Fig. 2(d) Precision and Recall measure is reported.

As we can see in Figg. 2 the results provided by iSoS are very encouraging, when the ontology is loaded we can see more relevant documents in the top 15 positions of the result set. This is confirmed by the Precision-Recall measure, the green curve shows higher performances then Google and iSoS without ontology.

We can conclude that using ontology knowledge for solving queries improves the relevance of the result set. This conclusion is mainly related to the words model discussed above, which can be used to analyze the content of documents and the meaning of words.



1	blog.wired.com/gadgets/2008/11/adobe-flash-on.html
2	www.techcrunch.com/2008/11/13/is-apple-building-a-search-engine/
3	www.sfgate.com/cgi-bin/blogs/sfgate/detail?blogid=19&entry_id=31899
4	**en.wikipedia.org/wiki/Apple**
5	oldcomputers.net/appleii.html
6	www.cnet.com.au/mobilephones/pdaphones/0,239036203,339279031,00.htm
7	www.cnet.com/apple-mac.html
8	www.mahalo.com/Apple
9	**www.botanical.com/botanical/mgmh/a/apple044.html**
10	news.cnet.com/8301-13506_3-10105263-17.html
11	news.cnet.com/8301-13579_3-10063765-37.html
12	apple20.blogs.fortune.cnn.com/2008/11/15/blackberry-storm-vs-apple-iphone-8-reasons-pro-and-con/
13	www.appleinsider.com/
14	apple2history.org/history/ah03.html
15	www.abcteach.com/directory/theme_units/apples/

(a)

r	URL	f	score
1	**en.wikipedia.org/wiki/Apple**	30	141.93791
2	**www.botanical.com/botanical/mgmh/a/apple044.html**	30	110.88
3	**www.hort.purdue.edu/newcrop/maia/history.html**	24	48.70992
4	**www.uga.edu/fruit/apple.html**	22	33.704483
5	**www.allaboutapples.com/**	18	48.295235
6	www.officesnapshots.com/category/apple/	18	18.988712
7	**www.applejournal.com/**	16	33.241737
8	www.cnet.com.au/mobilephones/pdaphones/279031,00.htm	14	123.47088
9	**www.fruitsandveggiesmatter.gov/month/apple.html**	14	29.38885
10	**www.applehill.com/**	12	36.22446
11	www.enchantedlearning.com/themes/apple.shtml	10	56.30069
12	blog.wired.com/gadgets/2008/11/adobe-flash-on.html	8	284.25266
13	oldcomputers.net/appleii.html	8	127.97332
14	www.cnet.com/apple-mac.html	8	116.25369
15	news.cnet.com/8301-13506_3-10105263-17.html	8	89.25321

(b)

1	www.apple.com/
2	en.wikipedia.org/wiki/Apple_Computer
3	www.apple.com/store/
4	www.info.apple.com/
5	developer.apple.com/
6	www.info.apple.com/
7	apple.slashdot.org/
8	**en.wikipedia.org/wiki/Apple**
9	https://support.apple.com/ibook_powerbook/batteryexchange/
10	finance.yahoo.com/q?s=Aapl
11	store.apple.com/uk
12	gizmodo.com/tag/apple/
13	discussions.apple.com/
14	developer.apple.com/iPhone/program/
15	support.apple.com/kb/HT1222

(c)

(d)

Fig. 2. Apple domain, query *apple*. 2(a) *iSoS* result without ontology. 2(b) *iSoS* result with ontology. 2(a) Google results. 2(d) Precision-Recall.

6 Conclusions and Future Works

We presented the main ideas behind *in Search of Semantics* project. At this time just results on light semantics computation have been discussed. A real environment, namely a web search engine has been developed in order to experiencing the methods for semantics representation. Since some experimental results are encouraging we can affirm that the main idea of this project must be still pursued. As future work we are preparing a full semantics environments for page ranking and we are working on methods for reveal and represents deep semantics.

References

1. T. L. Griffiths, M. Steyvers, J.B.T.: Topics in semantic representation. Psychological Review 114 (2007) 211–244
2. Colace, F., Santo, M.D., Napoletano, P.: A note on methodology for designing ontology management systems. In: AAAI Spring Symposium. (2008)
3. Santini, S.: Summa contra ontologiam. International journal on data semantics submitted (2007)
4. Derrida, J.: De la grammatologie. Paris:Minuit (1997)
5. Eco, U.: A theory of semiotics. Bloomington:Undiana University Press. (1979)
6. Ericsson, K.A., Kintsch, W.: Long-term working memory. Psychological Review. 102 (1995) 211–245
7. Kintsch, W.: The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review 95 (1988) 163–182

8. Potter, M.C.: Very short term conceptual memory. Memory & Cognition (1993) 156–161
9. Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior (1969) 240–247
10. Steyvers, M., Griffiths, T.L., Dennis, S.: Probabilistic inference in human semantic memory. Trends in Cognitive Science 10 (2006) 327–334
11. Chaitanya, C., Padhraic, S., Mark, S.: Combining concept hierarchies and statistical topic models. In: CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, New York, NY, USA, ACM (2008) 1469–1470
12. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. MIT Press (2004)
13. Santini, S., Gupta, A., Jain, R.: Emergent semantics through interaction in image databases. IEEE Transactions on Knowledge and Data engineering 13 (2001) 337–51
14. Grosky, W.I., Sreenath, D.V., Fotouhi, F.: Emergent semantics and the multimedia semantic web. In: SIGMOD Record. Volume 31. (2002) 54–58
15. Marr, D.: Vision. Freeman, S. Francisco,CA (1982)
16. Ballard, D., Brown, C.: Computer Vision. Prentice Hall, New York, N.Y. (1982)
17. Pylyshyn, Z.: Situating vision in the world. Trends in Cognitive Sciences 4 (2000) 197–207
18. Fortuna, B., Mladeni?, D., Grobelnik, M.: Semi-automatic Construction of Topic Ontologies. In: Semantics, Web and Mining. Springer Berlin / Heidelberg (2006)
19. Ding, Z., Peng, Y., Pan, R.: A bayesian approach to uncertainty modeling in owl ontology. In: Proceedings of the International Conference on Advances in Intelligent Systems - Theory and Applications. (2004)
20. Anderson, J.R.: The adaptive nature of human categorization. Psychological Review 98 (1991) 409–429
21. Roland G. Fryer, J., Jackson, M.O.: Categorical cognition: A psychological model of categories and identification in decision making. Working Paper Series National Bureau of Economic Research (2003)
22. Eco, U.: Kant and the Platypus: Essays on Language and Cognition. First Harvest edition (1997)
23. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003)