

The Case for Context-Dependent Dynamic Hierarchical Representations of Knowledge in Medical Informatics

Stefan V. Pantazi, MD, PhD¹, Isabelle Bichindaritz, PhD², Jochen R. Moehr, MD, PhD¹

¹School of Health Information Science, University of Victoria, Victoria, BC, Canada

²Institute of Technology/Computing and Software Systems, University of Washington, Tacoma, USA

Keywords: knowledge representation, information retrieval, case-based reasoning, unsupervised grammar induction

Abstract

We begin this article with an introduction of *associative concept spaces* and their properties and reiterate an earlier proposal for a definition of Medical Informatics. We focus on *dynamic representations of knowledge* and contrast them with ontological approaches that imply strong commitments to particular conceptualizations of reality. This focus leads naturally to *unsupervised machine learning* and *inductive learning* that include what is known in literature as *unsupervised classification* and *grammar induction*. Next, we provide an overview of hierarchical representations and propose a unifying view. Further, we review the task of grammar induction and by means of an example we show how the task of unsupervised classification is equivalent to a grammar induction procedure complemented by the use of *inverted indexes*. In the last section of the article we explore the implications of context-dependent, dynamic representations of knowledge for Medical Informatics and advocate an extension of the *concept of evidence in biomedicine*. Finally, we acknowledge the major hurdles implied by this proposal, which consist of increased complexity and of the need for privacy and confidentiality.

1 Introduction

In an earlier publication¹, associative concept spaces have been defined and their properties identified to be: high dimensionality, immense sparseness, dynamicity, and a similarity-based organization. There, the focus was on the *high dimensionality* and *sparseness* properties of concept spaces and led to defining Medical Informatics as context-dependent representation and processing of medical information. Though this definition aligned well with knowledge centric views that regard activities such as *ontology development* to be fundamental to Medical Informatics², some differences are also emerging. In this article we address this and shift the focus on the conceptual exploration of the *dynamicity* property of representations, a property which contradicts the *ontological commitment* process implied by ontology development activities. The commitment to particular understandings, theories, schemas and conceptualizations of reality, hurts representations for the simple reason that *uncertainty* is always a part of our reality and this leads to *incompleteness* of representations. To put it in another way, organizing knowledge in a universally applicable, context-independent, static, non-evolving conceptual structure (i.e., ontology) which one could commit to and apply to future problem solving, is bound to be problematic since one rarely possesses all the data in advance (i.e., attain completeness or be certain about the future). The recognition of this fundamental limitation leads naturally to the idea of *unsupervised machine learning*, particularly that of an *inductive learning* able to automatically reiterate and rebuild its theories dynamically (i.e., learn) in light of new data. In this case, the commitment to particular structures and the inherent representation errors are only temporary and the potential to overcome them reside in the learning capabilities of the system. As will be shown further, such learning strategies require hierarchical and dynamic context-dependent representational approaches. To further these points we place ourselves in the realm of natural language processing and follow with an overview of hierarchical representations and *grammar induction*.

2 Hierarchical representations

The importance of compositional and hierarchical approaches to both artificial and biological information processing is indisputable. Their universality transcends academic boundaries. In the context of the biological significance of self-organizing maps, Kohonen states it unequivocally³: “human brain is undoubtedly hierarchical.” The most important reason for using hierarchical and compositional approaches most likely lies in the inherent *information compression* capability of hierarchical models. Compositional hierarchies are also often described in cognitive science, psychology and memory research literature though rarely as an explicit method to overcome multidimensionality. One such example is the generic concept of “chunking”, a form of grammar induction that aims at the “meaningful packaging of information”⁴ and which usually shares discourses with concepts such as memory span, trace decay and the magic number seven (i.e., the average number of chunks that our short term memory can remember)⁵. Yet the most compelling examples of compositional and hierarchical models and representations seem to be those that come from the realm of language processing. Structurally, hierarchies pervade all aspects of natural language discourse where artifacts are commonly organized in genres, library

categories, volumes, individual books, chapters in books, sections in chapters, subsections, paragraphs, sentences, phrases, words and punctuation, morphemes, syllables, characters, phonemes, all part of different layers of the same hierarchy.

In the specific contexts of medical natural language processing and medical terminology development, notions such as hierarchy and compositionality are indispensable. Their application ranges from morphosemantic decomposition of compound medical terms⁶⁻¹¹ to the creation of compositional medical terminology schemes (e.g. GALEN¹² and SNOMED¹³) and to the compositional organization of anatomical concepts into hierarchical frameworks¹⁴. From the literature and existing applications, it is clear that there are two major types of hierarchical approaches to representation: simple hierarchies (total order structures) and multiple hierarchies (partial order structures). By taking into consideration some additional qualitative aspects of hierarchies, i.e., their supervised/unsupervised construction and fixed/dynamic nature, we arrived at a unifying view of hierarchical approaches to representation which can therefore be categorized as:

1. Simple classifications (mono-hierarchies)
 - a. Static mono-hierarchy
 - i. Supervised static mono-hierarchy (e.g., the table of contents of a textbook, International Classification of Diseases ICD9)
 - ii. Unsupervised static mono-hierarchy (e.g., hierarchies generated by cluster analysis and automated classification approaches)
 - b. Dynamic mono-hierarchy
 - i. Supervised dynamic mono-hierarchy (e.g., dynamic folder hierarchies, directory structures)
 - ii. Unsupervised dynamic mono-hierarchy (e.g., dynamic hierarchies generated by cluster analysis, conceptual clustering, and automated classification approaches)
2. Multiple classifications (poly-hierarchies)
 - a. Static poly-hierarchy
 - i. Supervised static poly-hierarchy (e.g., type hierarchies, formal concept analysis structures, ontologies, semantic networks, Medical Subject Headings - MeSH, WordNet)
 - ii. Unsupervised static poly-hierarchy (e.g., Latent Semantic Indexing, Self Organizing Map)
 - b. Dynamic poly-hierarchy
 - i. Supervised dynamic poly-hierarchy (e.g., manual document indexing based on MeSH in Information Retrieval, Case-Based Reasoning)
 - ii. Unsupervised dynamic poly-hierarchy (dynamic text categorization approaches, unsupervised indexing in Information Retrieval, Case-Based Reasoning).

According to this view, the most powerful hierarchical approaches to representation seem to be those that are dynamic poly-hierarchies (partial orders) constructed with minimal supervision. Information retrieval models, though rarely referred to as hierarchies because of their very dynamic nature, have the capacity to dynamically classify a document in multiple categories (i.e., retrieval sets) based on its content and similarity to a query (i.e., the category). This behavior makes them functionally equivalent to a poly-hierarchy whose dynamically created categories are the result sets of a query. Though significantly more difficult to build and maintain using current technology, poly-hierarchies are a well-recognized desideratum of medical terminology systems¹⁵.

Finally, hierarchical representations can, in fact, be regarded as formal grammars that have the capability to tremendously reduce the conceptual dimensionality of information by rewriting it in ways that discover and expose the internal structure of that information. From this perspective, any form of classification, be it manual, supervised or unsupervised, could be regarded as a form of grammar induction, a difficult information processing task which necessarily steps at a higher level of information processing complexity and becomes context-dependent in nature¹⁶. In order to support the equivalence of indexing in Information Retrieval with the task of grammar induction complemented by inverted poly-hierarchical structures we proceed with an overview of grammar induction followed by a proof of concept experiment.

3 Grammar induction

The specification, instructions, or the set or rules of how to construct something from its components is a *grammar*. A grammar can also be regarded as a theory or a model and applying it in order to construct something equates to a theory application. However, real world problems often pose the inverse of this problem, namely, theory elicitation or grammar induction, which consists of making up the specifications, instructions, grammar rules, from a given dataset. For example, in the case of computer strings, inducing a grammar is the basis to creating a more meaningful, potentially compressed packaging that changes some of the properties (e.g., dimensionality, sparseness) of the representation space in which the strings are represented. Though grammars can describe finite sets of objects, grammar induction is often regarded as “the identification of an infinite structure (i.e., a language) with a finite structural description (i.e., the grammar) on the basis of a finite number of examples”¹⁷ a definition which reiterates the idea that grammars are abstractions (i.e., theories). For natural

languages, grammar induction is also an extremely difficult problem, especially if one expects results to resemble the syntactic analysis derived by a linguist¹⁸. Even if one does not aim specifically at linguistically correct structures, there are theoretical proofs which prevent results that satisfy optimality criteria¹⁷. For example, the problem of deriving the smallest grammar – with obvious applications in data compression – is known to be NP complete. Similarly the possibility to identify a context free grammar has already been proven formally¹⁹ to be unattainable only from limited positive examples of an infinite language (i.e., identification in the limit). As a consequence, grammar induction approaches generally aim at approximate results, are guided by purpose (e.g., syntactic parsing, chunking, semantic disambiguation, etc.), and typically try to make use of any available apriori knowledge in order to improve results (i.e., supervised approaches). Despite its difficulties, the problem of grammar induction is an intensively studied topic in various contexts but predominantly in language acquisition domains¹⁷, hierarchical chunking²⁰⁻²² grammatical inference²³, unsupervised language acquisition²⁴.

3.1 “Days of week” experiment

For this proof of concept experiment, the “days of week” names have been input into the unsupervised grammar induction algorithm described in existing doctoral work¹⁶. The algorithm has detected existing patterns in the data and has created the compressed representation (i.e., grammar) of the seven strings shown in Table 1.

The patterns contained in the rewrite rules are features that could be used to dynamically search and organize the representation in context/content-dependent, meaningful categories. This capability is especially useful when the strings form sets of hundreds and thousands of items. As a simple example, given their formal grammar representation in Table 1, the way to categorize the days of week according to their content (and context) is to create an *inverted representation* which indexes on the features (e.g., **-nday**), **-sday**) and leads naturally to a multiple hierarchy showing the relationships between items and their features (e.g., **Sunday** and **-nday**) as well as between features themselves (e.g., **-sday**) and **-esday**) (Table 2).

This multiple hierarchy is identical to the general concepts of *inverted file*²⁵ and *inverted index*¹⁸ which are specific to information retrieval. As a consequence, the representations in Table 1 can be successfully used in a search and *retrieval on secondary keys*²⁵.

To summarize, the grammar induction procedure has allowed us to automatically and dynamically create a context/content-dependent categorization of the seven strings that name the seven days of week. According to this particular categorization, all strings are of type **-day**), some of which are of type **-nday**) and some **-sday**). Those of type **-sday**) can also be **-esdays**). Because some strings belong to other categories (e.g., those containing **-ur-**, **(T-** and **(S-**) the hierarchy is multiple and forms a partial order set¹⁶. This result is an example of

Table 1. Example of machine induced formal grammar of the seven strings that name the seven days of week; the rules that contain only terminals are grayed and the chunks in the rule expansions are explicitly delimited by | symbols

• Rewrite rule	• Rule expansion
• 4 → d a	• d a
• 3 → 4 y	• da y
• 2 → 3)	• day)
• 1 → n 2	• n day)
• 5 → (M o 1	• (M o nday)
• 6 → (S	• (S
• 7 → u r	• u r
• 8 → 6 a t 7 2	• (S a t ur day)
• 9 → 6 u 1	• (S u nday)
• 10 → (T	• (T
• 12 → s 2	• s day)
• 11 → e 12	• e sday)
• 13 → 10 u 11	• (T u esday)
• 14 → 10 h 7 12	• (T h ur sday)
• 15 → (W e d n 11	• (W e d n esday)
• 16 → (F r i 2	• (F r i day)

Table 2. Example of inverted (feature indexed) multiple hierarchy derived from the machine induced formal grammar in Table 1; the chunks corresponding to rules that contain only terminals (grayed) are the first level of the hierarchy

• (T-	10 → (T
• (Tuesday)	13 → 10 u 11
• (Thursday)	14 → 10 h 7 12
• (S-	6 → (S
• (Saturday)	8 → 6 a t 7 2
• (Sunday)	9 → 6 u 1
• -ur-	7 → u r
• (Sat <u>ur</u> day)	8 → 6 a t <u>7</u> 2
• (Th <u>ur</u> day)	14 → 10 h <u>7</u> 12
• -da-	4 → d a
• -day)	3 → 4 y, 2 → 3)
• (Friday)	16 → (F r i <u>2</u>
• (Saturday)	8 → 6 a t <u>7</u> 2
• -nday)	1 → n <u>2</u>
• (Monday)	5 → (M o <u>1</u>
• (Sunday)	9 → 6 u <u>1</u>
• -sday)	12 → s <u>2</u>
• (Thursday)	14 → 10 h 7 12
• -esday)	11 → e <u>12</u>
• (Tuesday)	13 → 10 u <u>11</u>
• (Wednesday)	15 → (W e d n <u>11</u>

an *unsupervised dynamic poly-hierarchy*, which, according to the classification in section 2 is one of the most powerful hierarchical approaches to representation and functionally equivalent to unsupervised indexing in information retrieval. The dynamic nature of this poly-hierarchy would be most evident in the case of a hypothetical need to reconsider the grammar in Table 1 as a result of the introduction of an eighth day of the week, which may be named, for example, Eightday.

4 Implications for biomedicine: redefining the concept of evidence

The implications for Medical Informatics are multiple. First, the proposed unifying view of hierarchical representations underlines the equivalence of *Information Retrieval* (IR) and *Case-Based Reasoning* (CBR), two paradigms whose mutual relevance has already been established²⁶. This is supported by the significance of dynamic hierarchical representations, of knowledge (particularly of *unsupervised dynamic poly-hierarchies*), and of inductive learning in CBR and IR. Second, given the importance of CBR and IR in biomedical research²⁷, we propose an extension of the definition of *biomedical evidence* to include knowledge in individual cases, suggesting that the mere collection of individual case facts should be regarded as evidence gathering. To support our proposal, we argue that the traditional, highly abstracted, hypothesis centric type of evidence that removes factual evidence present in individual cases, implies a strong *ontological commitment* to methodological and theoretical approaches, which is the source of the never-ending need for *current* and *best* evidence, while, at the same time, offering little provisions for the reuse of knowledge disposed of as obsolete. By contrast, the incremental factual evidence about individuals creates, once appropriately collected, a growing body of context-dependent evidence that can be reinterpreted and reused as many times as possible.

Currently, the concept of evidence most often refers to an abstract proposition derived from multiple, typically thousands of cases, in the context of what is known as a *randomized control trial*. Hypothesis forming is the cornerstone of this kind of biomedical research. Hypotheses that pass an appropriately selected statistical test become evidence. However, the process of hypothesis forming also implies a commitment to certain purposes (e.g., research, teaching, etc.), and inherently postulates ontological and conceptual reductions, orderings and relationships. All these are direct results of the particular conceptualizations of a researcher that is influenced by experience, native language, background, etc. As with any process of abstraction there is a high burden, and hence a high responsibility from the part of researchers to reduce, i.e., to discard only the information which is thought to be non- or less relevant for a certain finding. This reduction process will always be prone to errors as long as uncertainties are present in our reality. In addition, even though a hypothesis may be successfully verified statistically and may become evidence subsequently, its applicability will always be hindered by our inability to fully construe its complete meaning. This meaning is fully defined by the complete context where the hypothesis was formed and which include the data sources as well as the context of the researcher that formed the hypothesis. The application of evidence cannot be completely detached from the context where it was originally formed and tested.

But the most worrisome are the prospects that tested hypotheses that form biomedical evidence, once obsolete, add to an increasing body of knowledge that has little chances of being reused. The reality of this situation is demonstrated by the never-ending quest for a type of evidence that is “current” and “best”²⁸, and hence by the implicit recognition of the existence of an also growing body of evidence which is “not current” and “not best”. Therefore, much of the effort associated with the discovery of older findings is wasted. The emphasis on *recency* of findings underlies the assumption that new is better and that old is not as good. The degree of this status quo can be easily surmised from the impressive amount of resources dedicated to ensuring that biomedical evidence meets these two important criteria of quality.

The discussion about commitment to research designs, methodological choices, and research hypotheses led us to the proposal to extend the definition and the understanding of the concept of evidence in biomedicine and align it with an intuitively appealing and an important direction of research: *Case-Based Reasoning* (CBR). From this perspective, the concept of evidence, traditionally construed on the basis of knowledge applicable to populations, is evolved to a more complete, albeit more complex construct which emerges naturally from the attempt to understand, explain and manage unique, individual cases. This new perspective of the concept of evidence is surprisingly congruent with the current acceptance of the notion of evidence in forensic science for instance. Here, by evidence, one also means, besides general patterns and trends that apply generally to populations, the recognition of any spatio-temporal form (i.e., pattern, regularity) in the spatio-temporal context of a case (e.g., a hair, a fibre, a piece of clothing, a smell, a fluid spot, a sign of struggle, a finger print on a certain object, the reoccurrence of a certain event, etc.) and which may be relevant to the solution to that case. This new view where a body of evidence is incremental in nature and accumulates dynamically in form of facts about individual cases is a striking contrast with traditional definitions of biomedical evidence. In addition, case evidence, once appropriately collected, represents a history that can be reinterpreted and reused as many times as necessary. But most importantly, the kind of knowledge where the “what is”, i.e., case data, is regarded as evidence can be easily proven to be less sensitive to the issues of *recency* (i.e., current evidence) and *validity* (i.e., best evidence).

4.1 Current evidence – the recency factor

Let us take the recency factor into discussion and put it in the context of old but carefully documented records. It is evident that recency of evidence embedded in case data, though still of importance in our ever-changing world, has a lesser impact: old evidence, in form of meticulously documented cases, are often invaluable resources for long term, retrospective research, whose potential is limited only by the documentation detail of those old records. The knowledge embedded in detailed historic records of cases will always have a chance of being reinterpreted in the future, provided one has the incentive and the resources to do this. This is not what usually happens with highly abstracted evidence that is deemed obsolete or is potentially overridden by newer findings. Because of the recency factor, this kind of evidence loses its significance considerably and has little chances of being reinterpreted because the original data from which it was abstracted is usually no longer available.

4.2 Best evidence – the validity factor

Establishing validity of evidence is even more problematic as it implies comparison of highly abstracted representations created by people with different backgrounds, experiences, and from different data sets. Most importantly, the comparison is made without access to original records of cases. In order to make comparison possible, one needs to submit to a researcher's particular data sources, methods, experience, and knowledge elicitation abilities. This commitment also implies a certain amount of trust in the knowledge elicitation abilities of researchers, in the validity of their methods and of their data collection strategies.

On the other hand, having access to the original case records would allow verifying the validity of knowledge elicitation processes, as well as reinterpreting, if necessary, the data for arriving at potentially different conclusions. This means that even trivial findings may still have a chance to be turned into useful ones provided that they are backed up by detailed, rich records that can be reinterpreted, revised and relearned from. The commitment is therefore only to the data sources and to the data collection approaches, but not to particular research methods, or to particular researchers' experience and knowledge elicitation abilities. And with the advent of unsupervised, intelligent data acquisition from sensor networks that can automatically capture detailed case records, researchers may, in the near future, only need to orchestrate the analysis and interpretation of data analyses, as many times as necessary and/or possible. This ability to evolve biomedical knowledge with few concerns about validity into incrementally better evidence is a form of inductive learning that can be made possible through the dynamic representations that we referred to throughout this article. Finally, to a certain extent, this ability could be one that may grant academic identity and possibly contribute to a theoretical basis to the field of Medical Informatics by clearly distinguishing it from other more traditional areas of biomedical research, such as Epidemiology and Biostatistics. At the same time this ability is the source of some of the most important hurdles that need to be overcome in order to make it work: increased complexity and the need for privacy and confidentiality of records.

5 Summary and conclusions

We proposed that the fundamental property of *high dimensionality* of associative concept spaces can be handled naturally by hierarchical representation approaches which are compositional in nature. Hierarchical representations are universal and the structure of natural languages and of natural language artifacts reflects this very well. We further proposed a unification of hierarchical approaches as formal grammars that classifies them according to three axes, i.e., their supervised/unsupervised construction, total/partial order structure and their fixed/dynamic nature. We suggested that the most powerful approaches to representation are those that are *dynamic partial order sets* constructed with *minimal supervision* currently found in Information Retrieval and Case-Based Reasoning. By means of a proof of concept example, we have also shown how the process of unsupervised classification is equivalent to a grammar induction procedure complemented by the use of inverted indexes.

In exploring the implications of dynamic representations we have proposed an extension of the concept of evidence in biomedicine that can accommodate the informatics research. We have shown that factual evidence encoded in rich, detailed case records has the property to transcend spatio-temporal boundaries and retain validity and relevance in time and space. This is exactly the case in much of the exploratory, descriptive and qualitative informatics research. By carefully conducted interviews and detailed textual, visual, audio, representations of reality of observations, researchers are able to considerably restrict the range of possible explanations and alternative hypotheses and reduce uncertainty, while at the same time, staying as uncommitted as possible to particular conceptualizations, hypotheses and theories. This is in contrast to the situation of abstracted knowledge, and hypothesis driven research which, once obsolete, has little chances of being revised without access to original contexts. However, the access to contexts is also the source of the major challenges of informatics: increased complexity of managing context-dependent representations (acquisition, retrieval, adaptation, reuse) and the need of privacy and confidentiality.

References

- [1] S. V. Pantazi, A. Kushniruk, and J. R. Moehr, "The usability axiom of medical information systems," *International Journal of Medical Informatics*, vol. 75, pp. 829-839, 2006.
- [2] M. A. Musen, "Medical informatics: searching for underlying components," *Methods Inf Med*, vol. 41, pp. 12-19, 2002.
- [3] T. Kohonen, *Self-Organizing Maps*, vol. 30, 3rd ed. Berlin Heidelberg New York: Springer-Verlag, 2001.
- [4] R. Bourtochouladze, "How Many Memory Systems Are There?" in *Memories are made of this: how memory works in humans and animals, Maps of the mind*, S. Rose, Ed. New York: Columbia University Press, 2002, pp. viii, 199.
- [5] G. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *The Psychological Review*, vol. 63, pp. 81-97, 1956.
- [6] C. Lovis, R. Baud, P.-A. Michel, and J.-R. Scherrer, "Morphosemantems Decomposition and Semantic Representation to allow Fast and Efficient Natural Language Recognition," presented at AMIA Annual Fall Symposium, 1997.
- [7] R. H. Baud, A.-M. Rassinoux, P. Ruch, C. Lovis, and J.-R. Scherrer, "The Power and Limits of a Rule-based Morpho-Semantic Parser," presented at AMIA Annual Symposium. Proceedings., 1999.
- [8] S. Schulz and U. Hahn, "Morpheme-based, cross-lingual indexing for medical document retrieval," *International Journal of Medical Informatics*, pp. 87-99, 2000.
- [9] A.-M. Rassinoux, P. Ruch, R. H. Baud, and C. Lovis, "Semantic Handling of Medical Compound Words through Sound Analysis and Generation Processes," presented at Proc AMIA Symp, 2000.
- [10] U. Hahn, M. Honeck, M. Piotrowski, and S. Schulz, "Subword Segmentation - Levelling out Morphological Varieties for Medical Document Retrieval," presented at Proceedings of the AMIA Annual Symposium, 2001.
- [11] S. Schultz, M. Honeck, and U. Hahn, "Biomedical text retrieval in languages with a complex morphology," presented at Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, 2002.
- [12] A. Rector, A. Rossi, M. F. Consorti, and P. Zanstra, "Practical development of re-usable Terminologies: GALEN-IN-USE and the GALEN Organisation," *Int J Med Inf*, vol. 48, pp. 71-84, 1998.
- [13] K. Spackman, K. Campbell, and R. Cote, "SNOMED RT: A Reference Terminology for Health Care.," presented at Proceedings of the 1997 AMIA Annual Fall Symposium, Philadelphia, 1997.
- [14] P. Cerveri, M. Masseroli, and F. Pincioli, "Remote access to anatomical information: an integration between semantic knowledge and visual data," presented at Proc. AMIA Symp, 2000.
- [15] J. J. Cimino, "Desiderata for controlled medical vocabularies for twenty-first century," *Methods Inf. Med.*, vol. 37, pp. 394-403, 1998.
- [16] S. V. Pantazi, "A Deterministic Dynamic Associative Memory (DDAM) Model for Concept Space Representation," University of Victoria, Victoria, Dissertation 2006.
- [17] P. W. Adriaans and M. M. v. Zaanen, "Computational Grammar Induction for Linguists," *Grammars*, vol. 7, pp. 57-68, 2004.
- [18] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press, 1999.
- [19] M. E. Gold, "Language identification in the limit," *Information and Control*, vol. 10, pp. 447-474, 1967.
- [20] G. Wolff, "REFERENCES FOR "HIERARCHICAL CHUNKING"," R. J. Solomonoff, Ed.: (personal communication), 2004.
- [21] C. G. Nevill-Manning, "Inferring Sequential Structure," in *Department of Computer Science*, vol. Ph.D.: University of Waikato, New Zealand., 1996.
- [22] S. Edelman, Z. Solan, D. Horn, and E. Ruppin, "Learning Syntactic Constructions from Raw Corpora," presented at 29th Boston University Conference on Language Development, 2005.
- [23] J. L. Hutchens, "Natural language grammatical inference," vol. PhD: University of Western Australia, 1994.
- [24] Z. Solan, D. Horn, E. Ruppin, and S. Edelman, "Unsupervised context sensitive language acquisition from a large corpus," presented at Proc. 2003 Conf. on Neural Information Processing Systems (NIPS), 2004.
- [25] D. E. Knuth, "Retrieval on Secondary Keys," in *The art of computer programming: Sorting and Searching*, vol. 3, 2nd ed. Reading, Mass.: Addison-Wesley, 1997, pp. 392-559.
- [26] I. Bichindaritz, "Memory Organization As the Missing Link Between Case Based Reasoning and Information Retrieval in Biomedicine," *Computational Intelligence*, vol. 22, pp. 148-160, 2006.
- [27] I. Bichindaritz and C. Marling, "Case-Based Reasoning in the Health Sciences: What Next?" *Artificial Intelligence in Medicine, Special Issue on Case-Based Reasoning in the Health Sciences*, vol. 36, pp. 127-135, 2006.
- [28] B. Haynes, "Of studies, syntheses, synopses, and systems: the "4S" evolution of services for finding current best evidence," *Evidence-Based Medicine*, vol. 6, pp. 36-38, 2001.