Interpretation of Data and Identification of Information are Sine Qua Non for Modern Digital Information Services

Ying He, Ph.D. and Mehmet Kayaalp, M.D., Ph.D.
U.S. National Library of Medicine, Bethesda, Maryland 20894

Digital information is at the center of the new scientific endeavor and, if managed carefully, it may bridge scientists across disciplines. Scientists, clinicians, and the public need enabling tools to overcome barriers to communication of biomedical information. Biomedical information systems need to (1) interpret queries and the needs of users accurately, (2) identify, evaluate and combine all relevant information among a comprehensive set of sources, and (3) provide users the right information that they seek.

To meet these objectives, we are building a multifaceted information architecture that is capable of containing a comprehensive set of biomedical knowledge and expandable with new types of information (Kayaalp 2004). As part of this project, we continually investigate new modeling methods for intelligent data interpretation and information identification.

Conventional search engines such as PubMed®, Ovid®, and Google® are successful in providing documents of interest if user queries are simple and straightforward. They may, however, fail in providing the right information when user needs are more sophisticated, since these popular search engines usually rely on simplistic data models rather than full-fledged information models. They search character strings that are provided in user queries and return only the documents that contain those strings. In other words, these systems do not attempt to identify the information represented in user queries. A capable digital information system, on the other hand, ought to interpret user queries as well as target documents.

Consider the following hypothetical case: Conducting a retrospective study on a large pediatric population, a medical student observes that two seemingly independent conditions *alacrima* (deficient tear production) and *hypoadrenocorticism* (insufficient corticosteroid production by the adrenal cortex) were co-occurring in three cases. To find out whether such cases were reported in the literature, she searches Medline® abstracts using these terms as the keywords of her query. Conventional search engines of her choice return 0 hit, since no Medline entry contains both of her keywords.

If there were MeSH® terms synonymous to her keywords, she might have accessed a subset of relevant abstracts, since search engines usually perform adequately well in such cases. Note however that an exponentially increasing number of new terms are introduced to the biomedical literature everyday; thus, a more comprehensive set of external information sources ought to be utilized in order to detect information hidden in biomedical texts and user queries.

We developed a prototype to interpret user queries and identify pieces of information in biomedical text. For this example, our prototype system returned 46 Medline abstracts with perfect precision.

One important reason behind this success story is that our system uses 137 different external vocabulary sources (incl. MeSH) compiled in UMLS® Metathesaurus. Furthermore, it identifies *hypoadrenocorticism* as a condition that always occurs in 15 different diseases, syndromes, or conditions (e.g., Addison's disease, Allgrove syndrome, and hypoaldosteronism), some of which have additional etiologic subtypes. For *hypoadrenocorticism* alone, the system finds 242 different terms describing 28 biomedical entities with this condition.

All 46 Medline abstracts that are found relevant were about Allgrove syndrome (AS), in which the two conditions that our medical student is interested in always co-occur. The system, however, missed 38 abstracts, most of which discuss AS without mentioning the term *alacrima*. Those articles could be easily identified if our external sources were indicating (or our algorithm were intelligent enough to discover) that alacrima is an essential condition in AS.

The performance of our system improves as we better utilize lexical, syntactic, and conceptual information embedded in user queries and target documents. The effectiveness of our approach depends not only on our algorithms but also on external information sources—in this study, UMLS Knowledge Sources.

A comprehensive evaluation of the system performance will be conducted in order to asses the value of our approach objectively. The system discussed here is a prototype front-end of the muON project (Kayaalp 2004), of which theoretical underpinnings come from machine learning, probability theory, computational linguistics, and scientific ontologies.

Acknowledgements
This study was supported in part by the Oak Ridge Institute of Science and Education. Scientific computations and modeling were conducted on the Biowulf cluster at the National Institutes of Health.

Reference
Kayaalp, M. (2004) Modeling and Learning Methods. *A Report to the Board of Scientific Counselors*. Report No. LHNCBC-TR-2004-002. Bethesda, MD: Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine.