THE LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

A research division of the National Library of Medicine

TECHNICAL REPORT
LHNCBC-TR-2007-001

Separation of Data, Interpreters and Likelihood

March 2007

Mehmet Kayaalp, M.D., Ph.D.

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Abstract

One of the most important fallacies in information processing is the assumption that a particular interpretation of any given data is truth. Layers of complex information structures and belief systems are constructed on this foundation without questioning this assumption at each new layer. For example, only a small portion of information in the entire scientific literature is backed up with complete, currently available data from which the information has been deduced. For the rest, the data are no longer available; hence, there is no way to reinterpret the data and validate their conclusions. A sustainable model of biomedical data, information and knowledge requires a repository of objective data and metadata, associated with a set of software interpreters or interpretation protocols (algorithms). Outputs of interpreters become new sets of data to be interpreted by other interpreters. This model preserves not only abstractions and other underlying assumptions explicitly but also enables coexistence of differing interpretations, which may yield different conclusions on the same data.

Introduction

Information is a valuable commodity if it is right, but it may be quite dangerous or damaging if the misinterpretation of data gives a false impression of truth. How can we distinguish valuable information from a dangerous one? Is there a way to decrease the chance of facing false truths?

As the reader knows well, there is no free lunch, and one has to dig deep in order to separate fact from illusion, which frequently is based on honest but false assumptions. In order to accomplish this, one needs to have access to all facts, including data and metadata, which indicate the way the data are interpreted. Without an explicit account of data, metadata and the method of interpretation, success would be difficult, if not impossible.

The most basic requirement of a scalable (sustainable) model of information is that the underlying data and metadata should be free from implicit assumptions and unspecified interpretations. When data are tainted, there may be no way to reach a reliable conclusion. Even though the reliability of a conclusion depends very much on the soundness of interpretation, the quality of interpretation is secondary since flawed interpretations can be detected by thorough analysis and be replaced with sound interpretations as long as the data stay pure.

Data and Interpreters

One of the central concepts of this paper is data; thus, we need to be clear about what the term ‘data’ signifies. Anything and everything that we can register or record is data. The record itself is data, not its content. The content may be right or wrong, vague or exact, precise or imprecise, but data do not have those characteristics.

Here we consider only discrete data; that is, when provided to any two arbitrary (human or software) agents that are competent to process the data, they should be able to reproduce the data exactly. For example, if a short English sentence is presented in written format to two literate, native English speaking adults without any significant disabilities, they should be able to reproduce the same sentence with the same sequence of letters, white spaces and punctuations.

A sentence written on a paper has both discrete and analog qualities. The two adults in our example would interpret only the discrete portion of the data, namely letters, whitespaces, and punctuations and ignore the analog portion, namely all other qualities of the image of the sentence, such as font, font size, spacing between fonts, which are discrete in computers but their printout is analog to human eyes.

The process of reduction of the data that is implicitly performed by these individuals is a type of interpretation called abstraction. One may abstract details of data consciously in order to simplify the information model in hand, or unconsciously due to incompetence. In this example, the sentence was first printed on paper (transformation of digital data from discrete to analog) and then interpreted by individuals (transformation of data from analog to discrete).

When data are processed by an interpreter, the resulting output deviates from the original data. It would be guaranteed that the interpretation would not adversely affect the original data only if the principle of separation of data and interpreters is followed.

Emergence of Information

The first duty of an interpreter is to decode the data and elicit the represented information. Then it constructs new information based on information from the original data as well as from the external data that may be embedded in the algorithm of the interpreter. Finally, it encodes the new information in the new data.

Usually, nontrivial information is not a function of several data points; rather, information becomes of interest as it emerges through layers of data transformed by a sequence of interpretations (see Figure 1).

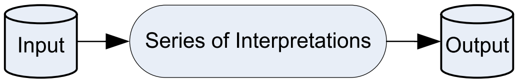

Figure 1. Conventional Process, where Interpreters, Metadata and Intermediary Data Are Fused together

Mostly, intelligent information processing involves a series of interpretations. A sample set of the interpretations involved in natural language understanding (NLU) is illustrated in Figure 2. Input is usually raw text (or speech in a wave form) and output may be a structured text with semantic labels. The set of interpreters and their order may differ in different NLU systems.

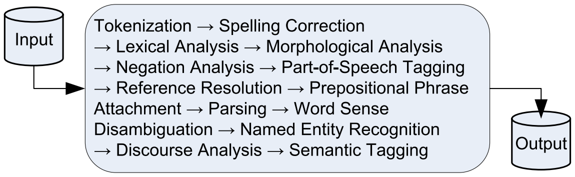
Input
Tokenization → Spelling Correction
→ Lexical Analysis → Morphological Analysis
→ Negation Analysis → Part-of-Speech Tagging
→ Reference Resolution → Prepositional Phrase
Attachment → Parsing → Word Sense
Disambiguation → Named Entity Recognition
→ Discourse Analysis → Semantic Tagging
Output

Figure 2. Sequence of Interpretations in NLU

Our cognitive processes such as thinking and decision making are not much different. However, due to our cognitive limitations, after reaching a new synthesis, which is a new state of understanding or abstraction based on our last interpretation, we discard the previous data from our mental models and keep the newly synthesized data instead. Discarding the processed data and keeping the abstracted information helps us ease our cognitive load and focus on new frontiers in order to further our understanding. Creation of scientific information also follows this conventional structure. Given that our cognitive abilities and limitations substantially differ from the speed and capacity of computers, imposing the modes of our mental processes on computing and information storage is neither necessary nor effective.

Even though it is not mentioned in Figure 2, most text-based information systems convert letters to lower-case format before doing anything else such as indexing. An index with all-lower-case data means substituting the original data with all-lower-case data (i.e., practically discarding the original data along with the case information). When these systems need to do named entity recognition or word sense disambiguation, they no longer have access to case-information of noun phrases such as *white house* or *turkey*.

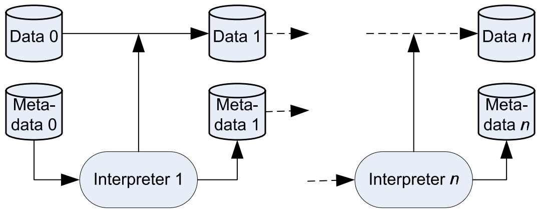

Figure 3. Traceable Transformation of Data after Following Principle of Separation of Data and Interpreters

The interest spectrum of users of biomedical information is so widespread that one cannot foresee all aspects of information at the point of data interpretation and information encoding. Therefore, as a rule of thumb, roads to alternative interpretations ought to be left open at every point of interpretation; we ought to encode every piece of information of possible interest and store the resulting data systematically so that we can effectively access them on demand (see Figure 3).

Facets—Alternative Views of the World

In the introduction, the following question was posed: Is there a way to decrease the chance of facing false truths? Preservation of data at the each step of interpretation is a way of decreasing that chance as it opens up to new opportunities for

- detecting and correcting interpreter errors,
- validating earlier analyses of data,
- updating earlier analyses with newer, advanced versions of interpreters,
- better understanding of interpretation processes by tracing data transformation,
- making metadata explicit,
- making interpretation assumptions and algorithmic information explicit, and
- reusing and sharing intermediary data.

Perhaps none of the above is as important as facets and multifaceted representation of information in the biomedical world. Facets constitute differing sets of data and metadata along with an associated interpreter, each of which represents a particular interpretation of the data.

Figure 4 illustrates that **Data A** created by **Interpreter A** is stored without discarding **Data**. The preservation of the original data indicates that the system does not assume that this particular interpretation is truth and leaves its door open to new interpretations. When **Interpreter B** is introduced, a new facet is opened and the new interpretation is encoded into **Data B**. As this example shows, facets of a system enable alternative interpretations of data coexist side-by-side.

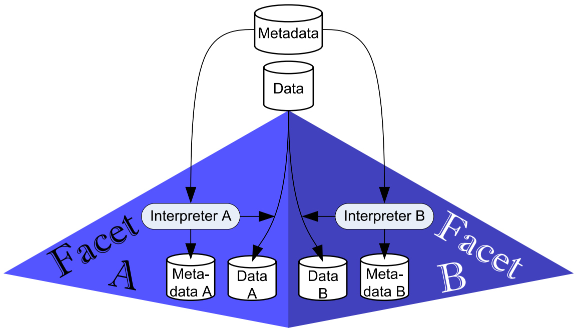

Figure 4. Representation of Different Interpretations of the Same Data in Facets

Each interpreter may act as a filter yielding a particular abstraction. Differences between facets may also be due to differing assumptions or different external data that interpreters may be using.

In the world we live, facets are everywhere. In an MRI study of the brain, different sequences of a particular slice are different facets, presenting different properties of the brain tissue on that slice. In a database management system, every table and every view is a facet. In a probability distribution, mean, median, mode, variance, skewness and kurtosis are different facets of the distribution. In the social realm, every issue has multiple facets, without which it is not possible to do a reasonably reliable analysis.

Although there are numerous facets that would be interesting to discuss here, a simpler example such as the tokenization of a term may provide a better insight on facets and their management. Tokenization is usually the starting point of every processes involving natural language processing and information retrieval. A token, which usually is a word, is a sequence of letters, numbers, and/or punctuations, and is considered as the unit of natural language information. Since it is usually performed before all other subprocesses (see Figure 2), errors and obscure algorithmic assumptions in tokenization are propagated without notice to all subsequent steps. If the resulting data are not stored, tokenization related problems that remained latent throughout the process may be very hard to detect and solve.

Tokenization is a difficult process, since algorithmic decisions and assumptions that are made in tokenization should always be remembered by the system designer during the rest of natural language processes. Due to high variance of styles and usages in natural language, a predictable tokenization of terms (especially newly introduced scientific terms such as gene names and biochemical terms) can be quite complex. The following simple example is just a glimpse of such difficulties.

We tested 13 tokenizers on MEDLINE® abstracts and saw a number of problems and inconsistencies among them [1]. An important class of inconsistency is on numeric terms such as 8.4-13.8%. Figure 5 shows 9 different outputs obtained from 13 tokenizers. In other words, the term is interpreted by 13 different interpreters yielding 9 different facets. In this example, tokenizers did not use any metadata nor produced any.

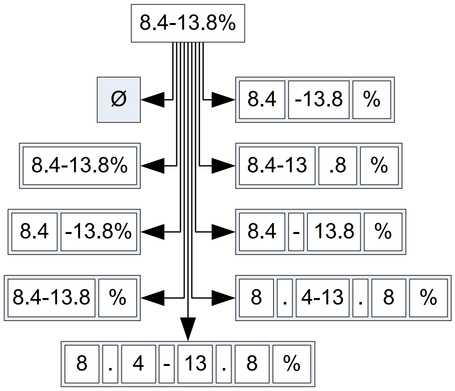

Figure 5. Nine Different Tokenizations of the Same Term

On terms that contain both letters and numbers (e.g., CO2, 12th, 5mg, 10S), the previous set of tokenizers yield interesting facets of tokens. Some keep them together within a single token, while others separate letters from numbers. Both options have pros and cons, and neither is necessarily better than the other, since some would be more useful for certain purposes, others for other purposes.

It is hard to find a perfect tokenizer that adapts to every style and provides outputs that are consistently in line with what we would expect. The ones that perform well for one example, do not necessarily perform as well for others. Using the union of all tokens obtained from different tokenization schemes may yield a more comprehensive index and better performing information retrieval.

Facets will be indispensable tools for information management, since different users and interpreters are interested in different facets of information. Abstraction is one of the most powerful tools that we and our software artifacts use. Repositories of the future that cannot organize different abstractions of the same data in a systematic fashion will be destined to fail in answering the needs of some users and will remain incompatible with a large number of software tools.

Emergence of Knowledge

In IT focused media, there is an overuse (perhaps to the extent of abuse) of the term knowledge. Usually the actual term used should be either data or information. In rare cases, where knowledge is the right term, one needs to find the answer to the following question: What properties of this system turns information into knowledge?

In the Section Emergence of Information, we argued that data become information when there is an interpreter that is capable of interpreting the data competently. In this view, information may be formulated as

Data + Interpretation = Information.

Information however is an intangible, ephemeral concept, similar to thought, and is shaped within the human interpreter's mind.[*] Up until the information is encoded by the interpreter into data, it stays obscure to the outside. We on the outside indirectly know that the interpreter identifies information in the data (i.e., that the information exists) based on the effect of information on the data transformation via the interpreter. The flow of information is realized as this process continues iteratively (see Figure 3).

We usually consider ourselves acquiring knowledge when we read papers or attend a seminar of interest. But according to the above model, we merely acquire information. Is something missing in the above model? Or, what is the difference between information and knowledge?

Let's analyze this issue by an example. Suppose we read the following sentence: *The Earth is flat.* The sentence certainly contains some information, since we can interpret it and understand what it says. On the other hand, unless we suspect any metaphor behind this line, we would immediately dismiss this information. In other words, to us, this sentence does not contain any knowledge.

[*] If the interpreter is not a human being but a software system then the mind would be a figurative term indicating a particular mechanism, such as finite state automata, that changes its state based on the data it receives.

Since Plato, classical epistemologists explain this phenomenon (or diagnose the problem) as falsity of the stated belief (i.e., information) [2]. On the other hand, we know that that sentence would not be dismissed as easily by our ancestors in antiquity and could have been considered knowledge. In the classical school, knowledge is associated with the absolute truth. The presence of uncertainty diminishes knowledge into belief. If we follow this line of reasoning, we should never say that we know something, given that none of us is the ultimate authority on truth. Even if there is no absolute truth, we still need the term and the concept of knowledge in order to be able to utter what *we* know and be able to differentiate *flat Earth* from *round Earth*. Although both statements contain information, we, in reality, assign them different truth values.

In classical logic, a statement has only one of two assignments, either true or false. Does this process (assigning a statement a binary truth value) yield knowledge?

Some, who may be characterized as Bayesians, believe that truth assignment is a subjective evaluation process [3]. It is subjective because given a piece of information, different people may have different degrees of confidence about its adherence to truth. The Bayesian solution is to assign a probability value to such statements, which, too, ranges from 0 (false) to 1 (true) but also covers everything in between.

The Bayesian approach seems more realistic and comprehensive than classical logic as it can define infinitely many shades of gray instead of just black and white; that is, the possibility that the information is true can be indicated with a probability value from an infinite set of probabilities on the real line between 0 and 1. In this model, knowledge emerges from information as its likelihood (i.e., the probability of the truthfulness of the information) nears to 1.0.

This model of knowledge is also compatible with the modern understanding of the nature of scientific knowledge. In this view, scientific knowledge is distinguished from dogma as it must be open to criticism and challenges by new evidences, through which it can be refuted.

In the Bayesian model, confidence (as the opposite of skepticism) in any particular scientific knowledge is represented by a likelihood parameter *p*, where $0 < p < 1$, which is the outcome value of a likelihood function.[*] In this view, knowledge may be formulated as follows:

Information + Likelihood = Knowledge

A facet at the knowledge level consists of not only data, metadata, and an interpreter, which are elements of a facet at the information level (see Figure 4), but also a likelihood parameter and an evaluator. Our second design principle, the **Separation of Information and Likelihood,** suggests that these new elements ought to be separated from the other elements of the facet at the information level. Here are some of the reasons behind this principle:

1. Different evaluators may assign different likelihood values to the same information.
2. As we accumulate more evidence about the information, we need to update the likelihood.
3. When likelihood is separated from information, the focus of criticism shifts from the represented knowledge to the value of likelihood, which is much easier to change than the represented information.
4. Changing the likelihood value at an early stage of information flow can be propagated to likelihood values of information at later stages through probability calculus. Propagation of

[*] Popper proposes a non-probabilistic version of the same notion and calls the resulting measure *truthlikeness* or *verisimilitude*.[7]

likelihood updates is much easier than altering the rest of information that are dependent on the information that had to be changed.

Discussions and Ongoing Work

Likelihood (or any other parameter of truthfulness) is a parameter about information; thus, we may call it a metainformation. Is likelihood the only metainformation? What constitutes a likelihood function? As there are a number of facets of information, there are a number of considerations in order to judge the truthfulness of information, such as coherence, rationality, reliability, completeness and precision, each of which can be considered a part of metainformation and should be factored into the likelihood function; however, these are beyond the scope of the paper. Discussions about coherence, rationality and reliability can be found in [3 4]. Due to scope of this paper, it is presumed that likelihood parameters are provided along with information. If the represented information is fully parameterized, any change of a particular parameter can be propagated to others through a chain of belief updates as done in other graphical probabilistic models such as Bayesian networks.

The principles that are discussed in this paper are the design principles of an actual system that is being developed in the Modeling and Learning Methods Project [5] at the Lister Hill National Center for Biomedical Communications at the US National Library of Medicine. The backbone of the system is called multifaceted ontological networks (muON). The current design of muON is based on a probabilistic graph structure where each node represents a particular piece of information and is related to other nodes probabilistically. The nodes contain any type of information, including biomedical concepts, linguistic concepts, linguistic structures and mathematical structures as well as actual data that are linked to the concepts and structures. The rationale behind the probabilistic nature of the design is discussed in [6]. The underlying probabilistic method called parameter interdependency networks (PIN) is also developed as part of the project. PIN models are graphical Bayesian models similar to Bayesian networks and chain graphs but with more extensive representation capacity and details.

The literature on information retrieval and information management systems is huge, but, to the best knowledge of this author, there exists no information retrieval, data warehouse or knowledge management system that adheres to the principles discussed in this paper.

The reason behind the lack of research on these design principles may be due to the intense focus on the performance of information systems and products. Compared to such research and development efforts, our aim is quite different. In line with the mission of our institute, we intend to capture biomedical data, information and knowledge to the largest extent possible, so that we can develop necessary information and knowledge tools to improve effective utilization of biomedical knowledge and to enable effective cross-disciplinary communication in biology, bioengineering, medicine and healthcare.

Adhering to the principles described in this paper yields a large amount of additional data, since processed data are not discarded at the end of each stage. However due to high redundancy and overlap between the datasets in subsequent stages of data transformation, a large amount of data can be reused and compressed.

Even though the paper discusses the rationale of the principles in depth, it does not discuss how the resulting data should be organized systematically, so that the effective use of data can be achieved. The organization for effective access and utilization is the task of muON and will be published in the future.

Acknowledgments

The author thanks to Dr. Clement J. McDonald and Dr. Susanne Humphrey for their thoughtful comments and suggestions on an earlier version of this report.

References

1. He Y, Kayaalp M. A comparison of 13 tokenizers on MEDLINE. Lister Hill National Center for Biomedical Communications, 2006; LHNCBC-TR-2006-003.
2. Bernecker S, Dretske FI, Eds. Knowledge: readings in contemporary epistemology. Oxford University Press, 2000.
3. Kyburg Jr. HE , Smokler HE. Studies in subjective probability. New York, NY: John Wiley & Sons, 1964.
4. Bovens L, Hartmann S. Bayesian epistemology. Oxford, UK: Oxford University Press, 2003.
5. Kayaalp M. Modeling and learning methods; a report to the board of scientific counselors. Bethesda, MD: Lister Hill National Center for Biomedical Communications, 2004; LHNCBC-TR-2004-002.
6. Kayaalp M. Why do we need probabilistic approaches to ontologies and the associated data? Proc AMIA Symp 2005.
7. Popper KR. Conjectures and refutations. 5th ed. London, UK: Routledge, 1989.