Communications of the IIMA

Volume 11 | Issue 1 Article 4

2011

From Data to Wisdom: The Progression of Computational Learning in Text Mining

Robert P. Schumaker
Cleveland State University

From Data to Wisdom: The Progression of Computational Learning in Text Mining

Robert P. Schumaker
Cleveland State University
USA
rob.schumaker@gmail.com

ABSTRACT

The DIKW hierarchy has long been a standard framework with which researchers can differentiate between levels of what they see and know. However much of the research conducted explores the nuances and precise divisions between each hierarchy level and assumes that the user will know how to use them. We plan to restrict our study to textual Web documents and propose a framework extension to the DIKW hierarchy that encompasses acquisition, delivery and prediction elements. We feel that such an extension can help better define each level of the DIKW hierarchy into discrete units that can be applied to the content contained within the Internet.

INTRODUCTION

> Knowledge signifies things known. Where there are no things known, there is no knowledge. Where there are no things to be known, there can be no knowledge. We have observed that every science, that is, every branch of knowledge, is compounded of certain facts, of which our sensations furnish the evidence. Where no such evidence is supplied, we are without data; we are without first premises; and when, without these, we attempt to build up a science, we do as those who raise edifices without foundations. And what do such builders construct? Castles in the air (Wright, 1829, p. 92).

These are the words of the Scottish-born Frances Wright in her 1829 book, *Course of Popular Lectures*. While Polanyi might argue that this description of knowledge signifying things that are known, ignores tacit knowledge (Polanyi, 1997), Wright's words of warning could easily be applied to all academic disciplines. This poetic linking together of knowledge, facts and data occurred over one-hundred and fifty years before Ackoff (1989) laid the foundations for the familiar data-information-knowledge-wisdom hierarchy (DIKW).

The DIKW hierarchy is widely accepted in knowledge management circles as a way to represent the different levels of what we see and what we know (Cleveland, 1982; Zeleny, 1987). With each successive level relying upon the previous, deeper levels of the hierarchy, this model provides an increasing awareness of the surroundings (Carlisle, 2006) where meaning can be found in the organization-wide continuum of data, information, knowledge and even wisdom (Chen, 2001). By correctly identifying and applying the DIKW framework, perhaps we can avoid Wright's "castles in the air."

The DIKW Framework

Figure 1 illustrates Alavi's linear interpretation of the DIKW Framework (Alavi & Leidner, 2001) and provides an example of how it functions.

Figure 1: DIKW Framework and Example.

While Alavi's portrayal of the DIKW framework as independent boxes arranged in a linear fashion has been the subject of debate (Fricke, 2008; Tuomi, 1999), the DIKW framework present in Figure 1 represents a more unifying view that each level is interwoven (Carlisle, 2006; Han & Chang, 2002; Hildreth & Kimble, 2002; Stenmark, 2002). However, it is still a commonly held view is that each level of the DIKW framework is derived from its immediate predecessor (Davenport & Prusak, 1998), just as its antithesis; misinformation, error, ignorance and stupidity is a compounding of prior levels as well (Bernstein, 2009). Below we define each element of the DIKW framework.

Data, the observable differences in physical states (Boisot & Canals, 2004), is acquired from stimuli and careful inspection of the world around us. To put it into Wright's perspective, we gather data from our senses, which is not entirely incorrect in our modern understanding as computers, lacking human senses, are still able to obtain data via external sources. Data by itself is generally overwhelming in volume and not entirely usable. In the example of Figure 1, data comes in the form of consumer grocery buying habits. In order to be of practical value, data must be transformed by identifying relationships (Barlas, Ginart, & Dorrity, 2005) or limited to only that which is relevant to the problem at hand (Carlisle, 2006). This transformation of content gives us Information.

If we were given these three grocery lists one at a time, it would be difficult to see the relations between them. An examination of just List 1; Apples, Bananas, Peppers and Pineapples, would lead a reader to note these are all fruits/vegetables and that they are arranged in alphabetical order. However, is this finding either important or relevant to a grocery store? Not necessarily. We find that there is not a sufficient mass of data to find anything worthwhile. Adding List 2 to the mix, we now have Avocados, Onions, Bananas and Apples. Now we can discount the alphabetic ordering assumption and still note the fruit/vegetable connection, however, it still might be too early to see a pattern in the data. Adding List 3 with Carrots, Bananas, Apples and Pumpkins should make it clear that Apples and Bananas occur often together, which provides us Information. While this is a simple and controlled example, real life grocery transactions may include dozens of items from thousands of unique shoppers, culminating in millions of visits. From this overwhelming amount of data, these otherwise unknown patterns are not so obvious and we must rely on computational tools to identify them. This is in essence the foundation of data mining.

While the value of Information may depend on its timeliness, accessibility, reliability and availability (Chen, 2005), it can be similarly argued that its value is also based on a particular user's need (Choo, 1996). In its essence, Information can be construed as meaningful, useful data (Bierly, Kessler, & Christensen, 2000). From the example of Figure 1, we gain Information from the observation that apples and bananas occur frequently together. Although this relation is not entirely useful at this stage, abstracting it to the next level of the hierarchy, Knowledge, can provide us additional meaning.

Knowledge is the aggregation of related Information (Barlas et al., 2005), that forms a set of expectations or rules (Boisot & Canals, 2004) which provides a clearer understanding of Information (Bierly et al., 2000). This level of the hierarchy begins the formation of rule-based systems which can allow individuals to expand their own knowledge while also benefiting the organization (Alavi & Leidner, 2001). In Figure 1, we can form an associative rule that links together apples and bananas. If I see Apples, I will also see Bananas.

The interweaving of data, information and knowledge permits the extrapolation of different levels in the hierarchy. Earlier instantiations of DIKW prohibited backwards movement (e.g., acquiring information from knowledge), however, modern research has questioned this assumption. Possessing knowledge can allow a user to derive information or even data (Stenmark, 2002) which stands in contrast to Alavi's original design. If there is knowledge that apples and bananas are linked together, then assumptions can be made regarding the composition of data. While these assumptions cannot recreate the data perfectly, there may be instances where data satisficing may be acceptable.

While the precise definitions of data, information and knowledge are still a matter of debate; wisdom can be viewed as a grasp of the overall situation (Barlas et al., 2005), that uses knowledge and knowledge alone (Carlisle, 2006) to achieve goals (Bierly et al., 2000; Hastie, Tibshirani, & Friedman, 2001). In Figure 1, wisdom can be depicted as the realization that increasing profits (our goal) can be obtained by cross-merchandizing two products that have a relation in consumer buying habits. Uncovering this truth rests in the capabilities of cognition

and human understanding (Carlisle, 2006), as a computational wisdom base is currently difficult to imagine Barlas et al., 2005). It is this incorporation of understanding that currently sets the divide between man and machine.

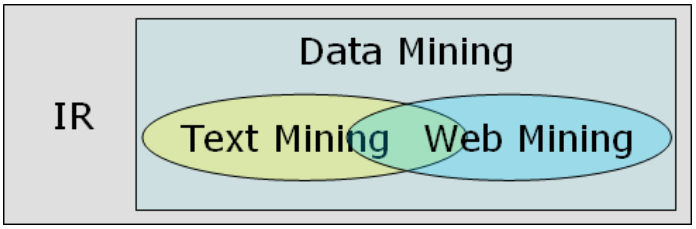
Textual Web Mining

With the advent of cheaper computer storage and interoperating networks, access to sources of data, information and knowledge has become readily available. Peter Drucker has characterized this period as the Knowledge Economy where the management of an organization's knowledge becomes a tool of competitive advantage (Drucker, 1969). Given the vast amount of content available, the Internet and especially the Web, is an abundant source from which we can extract valuable knowledge. Unfortunately, these sources are often unstructured and full of irrelevant material. While the act of finding data has become much easier, finding clean and well-organized data has been a challenge.

In order to address the difficulties of searching for relevant data, the science of information retrieval (IR) was created to sift through documents and databases to reduce information overload by returning those results that most closely match the query ("Information," 2007) which include search engines and question-answer systems. However, for the purposes of this paper, we will be focusing on a sub-area of information retrieval and data mining.

Data mining involves procedures to uncover hidden trends and develop new data and information from previous data sources. These sources can include well-structured, well-defined databases or the more common form of unstructured texts. While most Web-based communication is textual data, finding information or knowledge within these Web documents can be of strategic value to the knowledge-based organization and gives rise to the area of textual web mining. Figure 2 shows the relation between textual web mining and its parent disciplines.

Figure 2: The Intersection of Textual Web Mining.

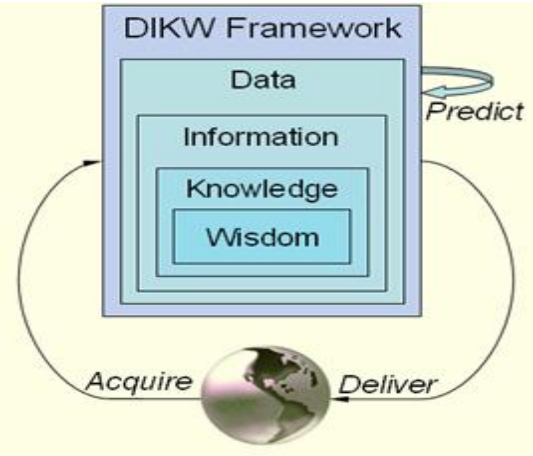


Textual web mining is the application of data mining techniques to extract knowledge from natural language sources of Web data, including Web documents, Web users and other Web data repositories (Hearst, 1999; Zhong, Liu, & Yao, 2002). This stream of research is a subset of Information Retrieval and uses Artificial Intelligence techniques to acquire and anticipate relevant information to meet specific domain needs.

Relating DIKW to Textual Web Mining

The exploration of the DIKW framework serves as an introduction to the topic of this paper; namely the focus on external interactions allowing the acquisition, delivery and prediction of textual Web content as shown in Figure 3.

Figure 3: Acquisition, Delivery and Prediction Framework.



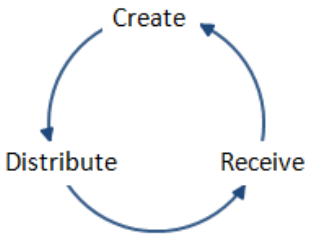
Acquisition can be defined as the process of obtaining relevant and concise content from Web sources. From Figure 3, content is acquired or transformed from external Web sources, such as Web documents, Web users and other Web resources.

Delivery is the process of returning relevant and concise content to Web users. Figure 3 illustrates the return of content back to the external environment. This process involves Information Retrieval tasks in order to identify the appropriate content requested.

Prediction is the process of projecting the trends and tendencies contained within existing content. In Figure 3, content can be used in the creation of further content. By leveraging existing textual Web sources and sound algorithms, content creation can become a valuable business tool.

This synthesis of Acquisition, Delivery and Prediction forms a circle of knowledge (Mazzotta, 1993) where Web users can create, distribute and receive content in a relevant and concise manner as shown in Figure 4.

Figure 4: Mazzotta's Circle of Knowledge.



Given Mazzotta's circle of knowledge, we could argue that knowledge distribution under Mazzotta would fit in our category of Delivery. However, Mazzotta's characterization of Create and Receive are not so straightforward when discussing textual Web documents and the roles computers play.

While we acknowledge that textual Web documents can be created and received by humans, we ask ourselves what role a computer system would play in this process. Computer systems can receive knowledge from textual Web documents, but what does it mean to create? We postulate that creation can be an imaginative/creative endeavor, as would be the general case in applying this to the novel works built by humans. However, computers lack this imaginative process but can nonetheless create new works or insights by forming associations, performing regressions and conducting probability calculations; all data mining operations. This implies that computers can predict new knowledge and we narrowly scope our prediction element to include the data mining aspects of Mazzotta's create category. Likewise, we also argue that not all knowledge that is created will be delivered. Certain knowledge may require an iterative creation process to further refine itself before it becomes useful. We posit that the create process should point back to itself, as we describe via the Predict function in our framework of Figure 3.

Further, we would point out that Mazzotta's circle of knowledge is for knowledge alone. We feel that this interpretation can be extended to include other levels of the DIKW hierarchy.

Acquisition

The process of acquisition is to transfer existing content and its structure into a computer-interpretable form (Potter, 2001). This content can come from humans or other sources such as textual documents or encyclopedias. When coupled with the Internet, content acquisition inherits new problems of scale such as information quality and reliability issues.

To focus on the higher levels of the DIKW hierarchy, knowledge acquisition has been a sought after goal since the early days of artificial intelligence. Newell posited that psychology and structure are important elements to perform a sequence of complex tasks and noted the similarities between cognitive tasks and existing programming languages. These languages are further engineered to use logic and conditional operators (Newell, 1973) to mimic human ability and to simulate human behavior (Feigenbaum & Simon, 1962).

Distribution

The Web is a vast distributed network of information. Users are constantly accessing and try to make sense of the Web's content using a variety of tools, such as search engines and digital libraries. The explosion of Web content volume coupled with the increasing ease of access to high speed bandwidth, means that researchers have a renewed focus on the design and implementation of content delivery platforms. In its simplest form, this may be a digital library where access tools facilitate the one-way flow of documents from the corpus to the end user. Another more dynamic approach is to allow the end users to be secondary contributors of information. This has been seen in electronic marketplaces of expertise such as Answer Garden (Ackerman, 1998; Ackerman & Malone, 1990) and the Annotate! system, which allowed

organizational workgroup-level document annotation to augment search engine results (Ginsburg & Kambil, 1999). In situations where all participants are potential information donors, coordination mechanisms are critical between the primary content authors and system administrators who are responsible for managing the knowledge-bases as they scale upwards.

Given a specific domain of interest and its audience pool, there are two important aspects of a networked knowledge transfer platform. We have knowledge delivery, where the system is able to answer a broad range of questions within the domain to the satisfaction of a broad range of the audience pool and knowledge acquisition, where the audience can contribute ideas to the system's knowledge base for the subsequent benefit of all.

Prediction

Acquiring relevant textual data is an important facet of prediction. While many textual Web documents are written on a daily basis, information flowing from these sources must be represented and transformed before existing applications can process it. This limitation forces open a temporal gap between when information is acquired to when it can be acted on. Information of an unexpected nature can cause a significant impact within its domain and the ability to harness these textual documents to make accurate predictions would be a useful decision-making tool.

Relevance to MIS Research

The decision-making process of incorporating diverse repositories of knowledge and managing it effectively is of paramount interest to decision-makers. While this knowledge may come from scattered domains, each with unique representational needs; it is necessary to integrate it within a unified framework that is flexible enough to the particular needs of the domain. Thus our findings of efficient and effective methods for leveraging such activities becomes of critical interest.

Our framework also complements the Design Science framework (Hevner, March, Park, & Ram, 2004) as a balance between the behavioral and technical aspects of MIS research. The acquisition and delivery components incorporate human-computer-interaction elements and computer mediated communication; both of which fit more on the behavioral side of this paradigm. On the other end of the spectrum, prediction is a wholly technical exercise. However, both behavioral and technical research combines to propose IT artifacts to extend the existing boundaries of management and enhance its effectiveness beyond what is already accepted.

CONCLUSIONS

The work presented in this paper highlights the process of acquiring, delivering and making predictions from textual Web documents within diverse domains. We have shown that textual Web content can be leveraged as a viable source for competitive advantage. Furthermore, this content can also be accurately represented, returned to the environment and built upon, all within the confines of the data-information-knowledge-wisdom (DIKW) framework. Overall, this paper

has outlined some of the domain-specific requirements needed for an effective and efficient acquisition, delivery and prediction mechanism. This work can be expanded over the next several years by looking further into the application of other techniques to the acquisition and delivery needs of other knowledge domains as well as tweaking existing methods in terms of knowledge prediction.

REFERENCES

Ackerman, M. (1998). Augmenting organizational memory: A field study of answer garden. *ACM Transactions on Information Systems, 16*(3), 203-224.

Ackerman, M., & Malone, T. W. (1990). Answer garden: A tool for growing organizational memory. In F. H. Lochovsky & R. B. Allen (Eds.), *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*, Cambridge, MA. New York: NY: ACM

Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16,* 3-9.

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Barlas, I., & Ginart, A., & Dorrity, J. L. (2005). Self-evolution in knowledge bases. In J. Romania & L. Batchler (Eds.), *Proceedings of IEEE Autotestcon '05 Conference* (pp. 325-331). Orlando, FL.

Bernstein, J. (2009). The data-information-knowledge-wisdom hierarchy and its antithesis. *Journal of Information Science, 35*(2), 68-75.

Bierly, P. E., Kessler, E. H., & Christensen, E. W. (2000). Organizational learning, knowledge and wisdom. *Journal of Organizational Change Management, 13*(6), 595-618.

Boisot, M., & Canals, A. (2004). Data, information and knowledge: Have we got it right? *Journal of Evolutionary Economics, 14*(1), 43-67.

Carlisle, J. P. (2006). Escaping the veil of Maya: Wisdom and the organization. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, Koloa Kauai, HI. doi: 10.1109/HICSS.2006.160

Chen, H. (2001). *Knowledge management systems: A text mining perspective*. Tucson: AZ: Knowledge Computer Corporation.

Chen, Y. (2005). Information valuation for information lifecycle management. *Second International Conference on Autonomic Computing* (ICAC'05), (pp. 135-146). Seattle, WA.

Choo, C. W. (1996). The knowing organization: How organizations use information to construct meaning, create knowledge, and make decisions. *International Journal of Information Management, 16*(5), 329-340.

Cleveland, H. (1982, December). Information as a resource. *The Futurist, 16*(6), 34-39.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business School Press.

Drucker, P. (1969). *The age of discontinuity: Guidelines to our changing society*. New York, NY: Harper and Row.

Feigenbaum, E. A., & Simon, H. A. (1962). Simulation of human verbal learning behavior. *Communications of the ACM, 5*(4), 223.

Fricke, M. (2008). The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science, 35*(2), 131-142.

Ginsburg, M., & Kambil, A. (1999). Annotate: A knowledge management support system. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Wailea, HI.

Han, J., & Chang, K C. -C. (2002). Data mining for web intelligence. *Computer, 35*(11), 54-60.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer-Verlag.

Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD. doi: 10.3115/1034678.1034679

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Hildreth, P. M., & Kimble, C. (2002). The duality of knowledge. *Information Research, 8*(1). Retrieved from http://informationr.net/ir/8-1/paper142.html

Information. (n.d.). In *Wikipedia*. Retrieval January 17, 2007, from http://en.wikipedia.org/wiki/Information

Mazzotta, G. (1993). *Dante's vision and the circle of knowledge*. Princeton, NJ: Princeton University Press.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing.* New York, NY: Academic Press.

Polanyi, M. (1997). The tacit dimension. In L. Prusak (Ed.), *Knowledge in Organizations*. Newton, MA: Butterworth-Heinemann.

Potter, S. (2003). A survey of knowledge acquisition from natural language. In *TMA of Knowledge Acquisitions from Natural Language*. Retrieved from http://www.aiai.ed.ac.uk/project/akt/work/stephenp/TMA%20of%20KAfromNL.pdf

Stenmark, D. (2002). Information vs. knowledge: The role of intranets in knowledge management. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences,* Waikoloa, HI.

Tuomi, I. (1999). Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems, 16*(3), 107-121.

Wright, F. (1829). Lecture IV: Religion. *Course of Popular Lectures* (pp. 85-105). New York, NY: G. W. & A. J. Matsell.

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management, 7*(1), 59-70.

Zhong, N., Liu, J., & Yao, Y. (2002). In search of the wisdom web. *Computer, 35*(11), 27-31.