



Framework for Interrogative Knowledge Identification

Fatimah Sidi (Corresponding author)

Computer Science Department, Faculty Computer Science & Information Technology

University Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

Tel: 60-12-203-8131 E-mail: fatimah@putra.upm.edu.my

Marzanah A. Jabar, Mohd Hasan Selamat, Abdul Azim Abd Ghani & Md Nasir Sulaiman

Faculty Computer Science & Information Technology, University Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

Tel: 60-3-8946-6555 E-mail: {marzanah, hasan, azim, nasir}@fsktm.upm.edu.my

Abstract

The difficulty of defining and capitalizing the knowledge in an organization from the business data captured in text files. These text files defined as unstructured document that is without a specific format example, plain text. Hence, this paper presents an Interrogative Knowledge Identification framework to identify unstructured documents that encompassed knowledge, information, and data. It tries to identify some high-level problems of the area from a higher perspective and then propose a possible solution thru the description of the framework. This research is an experimental approach using an appropriate test collection of unstructured documents. A system was developed based on the Interrogative Knowledge Identification framework. The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics compared with an expert. This is to improve better understanding the process of making sense the information or knowledge residing in unstructured documents.

Keywords: Knowledge identification, Interrogative, Unstructured documents

1. Introduction

The difficulty of defining knowledge in unstructured documents is due to the paradox that knowledge resides in a person's mind and at the same time, it has to be captured, stored, and reported. For that, philosophers classify knowledge into knowing-that and knowing-how. Knowing-that is factual where data are stored in databases and facts can be recalled, processed, and disseminated. While knowing-how is actionable to do something, turning data into information and in turn into knowledge (Spiegler, 2003).

It is estimated that 90% of electronically available material is unstructured and the amount of unstructured textual documents, accessible through the web, intranets, news groups, etc. is enormously increased every year (Iiritano & Ruffolo, 2001). Hence, huge amount of unstructured documents are available on the web and intranets. The amount of information available to us is constantly increasing and our ability to absorb and process this information remains constant. Apparently, knowledge exists and is found everywhere (ubiquitous) in unstructured documents (Feldman, 1999), so identifying and extracting knowledge in unstructured documents is essential.

2. Unstructured Documents

A document is a paper or set of papers with written or printed information, especially of an official type. It is categorized into two classes, unstructured and structured. Unstructured document (a "flat" document) will not have any attributes. These types of documents usually have a title, but after that the content is not organized in any structured fashion examples news and scientific papers.

Structured documents have a well-defined hierarchical structure, such as titles and sections clearly marked with single or multiple level headings. Other attributes that create hierarchy, such as distinctive colour, underlines, boldness, etc., are also considered. Structured form/scheme is the way in which data or information are arranged or organized in rows

and columns.

Unstructured documents cannot be queried in simple ways. Therefore, knowledge contained in unstructured documents can neither be used by automatic systems nor could be understood easily and clearly by humans. Hence, identifying knowledge from unstructured documents to be easily realized and understood by humans is one of the most valuable areas to be explored.

3. Spectrum of Data, Information and Knowledge

There are three theories of knowledge (As-Sadr, 1987; Cornford, 1957). First, the idealistic notion like Plato believes that knowledge was a function of the recollection of previous information. Second, the materialistic notion that believes in five senses. They consider sense perception as the source or means of knowledge. Third, the Islamic notion that believes in the existence of matter as well as soul. By that, knowledge is a complex concept and it is not easy to define because it is not easily understood, perceived, and measured. It has absolute truth, or ground truth, which describes the rich truths of real situation experiences (Davenport & Prusak, 2000; Drucker, 2001).

However, most people have some understanding of what knowledge is. Knowledge, information, and data are not interchangeable concepts. A brief comparison of data, information, and knowledge based on literature are tabulated in Table 1. It shows that data, in and of itself is a symbol, are out of context and with no value until processed into useful forms. By adding meaning, values, and searching for context to make sense of data, this context reveals the structure or relationship (or both) that organizes the data into information. Knowledge is the process of making sense of information. Examples of knowledge are patents, recipes, formulas, instructions, and designs. Without the dimension of context, culture, tacit, and time, knowledge will be little more than information. Thus, knowledge has more to do with who is interpreting the information (their own principles and values) than the objective information on which it is based.

3. Theoretical Background

This section discusses theoretical foundation of the framework proposed. Philosophers see knowledge as justified true belief, while scientists see knowledge as documents empirical research, supported by Quigley and Debons, (1999). The word data, information, and knowledge have many meanings in many contexts, which are often embedded in documents, repositories, processes, practices, and norms. Therefore as a foundation of this research, the basis approach of knowledge understanding adopted is the scientist views. While, on the knowledge management (KM) understanding, it is concerned on the knowledge growth in an organization (Steels, 1993), such as organizing of knowledge (Gurteen, 1999; MingYu, 2002). Hence, it is important to facilitate knowledge transfer or discovery in unstructured documents.

Unstructured documents are stored at any time in history. These texts are stored in hardware and retrieved through software, which contains data, information, and knowledge, each with its own characteristics and value. According to Quigley and Debons (1999), a cognitive spectrum of data, information, and knowledge focus on data-as-thing, information-as-thing and knowledge-as-thing located within text strings. They reported an interrogative-based approach to differentiate and quantify information and knowledge within text. The interrogative-based approach is described as the “who, when, what, where, how and why” analysis. Analysis using interrogative theory makes distinctions between data, information, and knowledge as follows:

- Knowledge text that answers how/why in the problem space
- Information text that answers when/where/who/what in the problem space
- Data text that answers no question in the problem space

They reported their finding that parsing of the paragraph into interrogative strings yields consistent results and a quantification of information and knowledge within the text. Based on the perspectives above, that data, information, and knowledge focus as-thing located within text strings. It is recommended that elements of personal components and dimension of context, culture, tacit, and time should be included in the discussion of the Interrogative Theory. This is because there is a lack of fluid mix of framed experience, values, contextual information, and expert insight in the spectrum of data, information, and knowledge. Values and beliefs are integral to knowledge, determine a large part of what the knower sees, absorbs and concludes from his observations. People with different values “see” different things in the same situation and organize their knowledge by their values. By that, challenges to incorporate the personal components of values and beliefs could be seen as the gap in the discussion of the Interrogative Theory as this theory sees data, information, and knowledge only as-things. Therefore, a new perspective of looking upon the spectrum of data, information, and knowledge can be derived by unifying Interrogative Theory and personal components of values and beliefs.

4. Interrogative Knowledge Identification Framework

The interrogative knowledge identification is used to address the need for the mechanism to identify knowledge from unstructured document in order to extract them. Briefly restating the interrogative knowledge identification, it identifies the type of document by separation of text into knowledge, information or data and unifying it with personal

components of values and beliefs. The approach of answering interrogatively is used to answer the question within the text in unstructured document to identify knowledge.

Another important aspect is to understand the process of making sense the information that resides in the unstructured documents into knowledge. Knowledge must have enough characteristics of information in terms of its meaning, values and context to reveal its structure or relationship or both. Lack of ways or methods to organize information of unstructured document would produce different knowledge from the same piece of information in different brains. Barachini (2003) reports that the more contexts stored with a chunk of information, the better the interpretation and transfer of knowledge. Hence, introducing interrogative contextual information by adding more contexts to the information, organizing, and structuring them into interrogatively structured form will increase better understanding and interpretation of knowledge that resides in unstructured document.

The interrogative contextual information is derived from the incorporation of context and additional information annotation with context key facility. Context is an abstraction of the context factors, which are represented as concepts (Schilit & Theimer, 1994). It is further exploited by Lamming and Newman (1992) as contextual information, where information entered into the computer is tagged with context keys facilitating future retrieval by using those keys. It is any information that can be used to characterize the situation of an entity (Abowd, Dey, & Abowd, 1999). An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. For that, the interrogative contextual information is utilized to understand the process of making sense of information into knowledge and maintain the meaning of the information. This is to gain the interpretation of the identical knowledge by classifying the main point of the unstructured document interrogatively.

The unification of the interrogative knowledge identification and contextual information with incorporation of personal components of values and beliefs is illustrated in Figure 1. The incorporation of personal components is motivated by looking at gaps and contradictions existing in retrieving documents through Internet by different people, culture, and values. Hence, the issue addresses is: how to identify knowledge in order to extract them in unstructured document? Different knowledge is produced from the same piece of information in different brains. Rationalization of the incorporation of personal components towards the interrogative knowledge identification is as follows:

- Nonaka (1994), personal components have a powerful impact on organizational knowledge;
- Davenport and Prusak (2000), knowledge is a fluid mix of frame experience, values, contextual information, and expert insight. It provides a framework for evaluating information. It originates in the mind of the knower to determine a large part of what the knower sees, absorbs, and concludes from his observations;
- Barachini (2003), knowledge is a private and personal thing. It is intuitive and strongly linked to the user's values and beliefs; and
- Virk (2004), manually transforming documents. Values are embedded because humans read documents, extract the values of existing fields, and then enter the values into a user interface.

The unification of the interrogative knowledge identification and contextual information with incorporation of personal components of values and beliefs as depicted in Figure 1 provides a proposal to establish an approach on transformation of extracted knowledge in unstructured documents by identifying, organizing, and structuring them into interrogative structured form. It is used to transform information in unstructured documents into knowledge. It is also used to understand the process of making sense of information into knowledge; maintain the meaning of the information; and gain the interpretation of the identical knowledge by classifying the main point of the unstructured documents interrogatively. It is designed to ease the burden of work, through augmentation and automation, allowing resources to be applied efficiently to the tasks for which they are most suited. It is important to note that not all knowledge extractor are computer-based, as paper and pen can certainly be utilized to generate, codify, and transfer knowledge (Ruggles, 1997). For the purpose of this research, however, the tools covered are primarily the technological ones due to their quick evolution, dynamic capabilities, and organizational impacts.

5. Research Setting

The research setting involves the development of the system based on architecture of the proposed framework; i.e. Malay/IK-Ontology (Malay Interrogative Knowledge Ontology). Basically, the system consists of these four processes:

- i. Prepare the unstructured documents to be processed and converted it into extension of plain text file.
- ii. Invoke lexicon identifier that uses lexicon interrogative analysis matching rules. It is used to identify and extract knowledge in each of the complete sentences written in the unstructured document. It is also used to extract interrogative lexical constructs from the individual unstructured document.
- iii. Invoke object recognizer that uses matching rules of object interrogative analysis to extract ontological constructs from the interrogative lexical constructs. It is used to populate objects and map the objects with ontology engineering. It

is a mechanism of a knowledge structure to represent the concept and relationship of the abstract model on how people think about things in the world.

iv. Transform ontological constructs to populate database scheme by connecting ontology model with conceptual modeling of object-relationship model. This is used to structure the extracted knowledge into interrogative structured form.

From the above processes, it can be simplified as shown in Figure 2.

This research is an experimental approach research using the Malay language. Therefore, an appropriate Malay test collection of Malay unstructured documents is required. Different topics are drawn such as main news, technology, editorial columns, sports, letters, and e-mails, while texts from children story books, articles, and magazines are drawn from Internet or retyped from the printed materials. This is a stratified population of data samples. In order to guarantee equal representation of each identified strata, a stratified random sampling is used. It is based on the number of words in the unstructured document and text which cover simple sentences constructed in Malay language. Each document drawn is assigned with a serial number and number of words.

The documents drawn are grouped according to the source of documents and range of number of words. The points of the range are defined at positions of 50-150, 151-300, 301-500, and the final range is more than 500. For each range, five unstructured documents are selected and sorted in ascending order by total number of words. The total number of words needed for Malay unstructured documents test collection is about 15% of 42,733 words from Malay Interrogative Knowledge Corpus (MalayIK-Corpus).

The Malay language corpus is derived from 6,000 word entries (about 4,000 root words and 2,000 derivations), a Malay language dictionary of Kamus Dewan published by Dewan Bahasa Perpustakaan (2005). It is also derived from other dictionaries of Kamus Imbuhan Bahasa Melayu (Ali, Shariff, & Dewa, 1993), Kamus Dwibahasa Oxford Fajar (Hawkins, 2001), and Kamus Komprehensif Bahasa Melayu (Othman, 2005). The sample used in this experiment, 15% of 42,733 words from MalayIK-Corpus are sufficient and justified to produce better results in extracting identified knowledge. It is more than the suggested by Gay and Airasian (2003, page 113) for sample of more than 5,000 units, a sample size of 400 (8%) should be adequate.

The results obtained are measured in terms of percentage of quantitative retrieval performance recall and precision metrics (Baeza-Yates & Ribeiro-Neto, 1999). The accuracy of the knowledge extracted is measured by precision (fraction of the retrieved knowledge which has been relevant), and recall (fraction of the relevant knowledge which has been retrieved). Comparison of results on the testing and analysis of the MalayIK-Ontology implemented is done with an expert evaluation. The Malay unstructured documents collection is given to the expert to identify the knowledge that resides in the collection interrogatively. The expert then validates the system generated output based on the interrogative criteria. The expert referred to is Prof. Dr. Hj. Awang Sariyan from Academy of Malay Studies, Universiti Malaya. He is also a member of the Language Committee Organizer Board, Institute of Language and Literature, Malaysia.

6. Results and Discussion

The analysis of results confirmed by a significant accuracy in identifying and extracting knowledge by using interrogative element of why. Unfortunately, this is not true for the interrogative element of how. Both these interrogative elements are used to identify knowledge within the text in unstructured document. Moreover, the analysis of results has also confirmed significant accuracy in identifying and extracting information for the interrogative elements of what and who. Unfortunately, the accuracy differences are not significant for the interrogative elements of where and when. The reasons for the performances differences are possibly caused by the quality of various formats and styles of writing the Malay unstructured documents collection used.

7. Conclusion

The paper presents a development of a system based on architecture of the Interrogative Knowledge Identification framework to identify unstructured documents that encompassed knowledge, information, and data. It also improved better understanding the process of making sense of information into knowledge. This framework can be used to organize and structure knowledge and information into interrogatively structured form which increased better understanding and interpretation of knowledge that resides in unstructured document. It shows a clear knowledge organization and structuring concept that can increase understanding of the concept among the community. This leads to potential increase sharable and reusable of the concept among the community. Moreover, it can be used to facilitate student learning in understanding the information and knowledge resides in the unstructured document.

Our future work is to enable the incorporation of personal components of values and beliefs integrate and contextual information within the proposed framework. This is to maintain the meaning of the information and gaining the interpretation of the identical knowledge in unstructured document which facilitate identical knowledge perceived by different people.

References

- Abowd, G. D., Dey, A. K., & Abowd, G. D. (1999, 27-29 September). Towards a Better Understanding of Context and Context-Awareness. *Paper presented at the Proceeding 1st International Symposium on Handheld and Ubiquitous Computing (HUC '99)*, Karlsruhe, Germany.
- Ali, H. M., Shariff, M. N. M., & Dewa, W. M. W. (1993). *Kamus Imbuhan Bahasa Melayu Edisi Kedua*. Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.
- As-Sadr, A. M. B. (1987). *Our Philosophy*. USA: Routledge and Taylor & Francis Group.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: Addison Wesley.
- Barachini, F. (2003). Frontiers for the Codification of Knowledge. *Journal of Information & Knowledge Management*, 2(1), 41-45.
- Cornford, P. F. M. (1957). *Plato's Theory of Knowledge*. Indianapolis, Indiana: Bobbs-Merrill Company, Inc.
- Davenport, T. H., & Prusak, L. (2000). *Working Knowledge: How Organizations Manage What They Know*: Harvard Business School Press.
- Dewan Bahasa Perpustakaan. (2005). *Kamus Dewan Edisi Keempat*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.
- Drucker, P. F. (2001). *The essential Drucker: Selections from the management works of Peter F. Drucker*. New York: Harper Business.
- Feldman, R. (1999, August). Mining unstructured data. *Paper presented at the Tutorial notes of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Gay, L. R., & Airasian, P. (2003). *Educational Research: Competencies for Analysis and Application*, Seventh Edition. Merrill, New Jersey: Upper Saddle River.
- Gurteen, D. (1999, February). Creating a Knowledge Sharing Culture. *Knowledge Management Magazine* 2.
- Hawkins, J. M. (2001). *Kamus Dwibahasa Oxford Fajar Edisi Ketiga*. Kuala Lumpur, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.
- Iiritano, S., & Ruffolo, M. (2001, 3-7 September). Managing the knowledge contained in electronic documents: a clustering method for text mining. *Paper presented at the Proceeding of the 12th International Workshop on Database and Expert Systems Applications*.
- Kaipa, P. (2000). Knowledge architecture for the twenty-first century. *Behaviour & Information Technology*, 19(3), 153-161.
- Lamming, M. G., & Newman, W. M. (1992). Activity-based information retrieval: Technology in support of personal memory. *Paper presented at the Proceeding of the IFIP 12th World Computer Congress on Personal Computers and Intelligent Systems - Information Processing '92*.
- MingYu, C. (2002). Socialising Knowledge Management: The Influence Of The Opinion Leader. *Journal of Knowledge Management Practice*.
- Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, 5(1), 14-37.
- Othman, A. (2005). *Kamus Komprehensif Bahasa Melayu*. Selangor Darul Ehsan, Malaysia: Penerbit Fajar Bakti Sdn. Bhd., a subsidiary of Oxford University Press.
- Quigley, E. J., & Debons, A. (1999). Interrogative theory of information and knowledge. *Paper presented at the Proceeding of the 1999 ACM SIGCPR conference on Computer personnel research*, New Orleans, Louisiana, United States.
- Ruggles, R. (1997). *Knowledge Tools: Using Technology to Manage Knowledge Better*. Paper presented at the *Innovation Working Paper*, Ernst & Young Center for Business Innovation in Cambridge, Mass., and Business Intelligence Ltd. .
- Schilit, B. N., & Theimer, M. M. (1994). Disseminating Active Map Information to Mobile Hosts. *IEEE Network*, 8(5), 22-32.
- Spiegler, I. (2000). Knowledge Management: A New Idea or a Recycle Concept? *Communications of Association for Information Systems*, 3(14).
- Spiegler, I. (2003). Technology and knowledge: bridging a "generating" gap. *Information & Management*, 40(6), 533-539.
- Steels, L. (1993). Corporate knowledge management. *Paper presented at the Proceeding of ISMICK'93*, Compiègne,

France.

Virk, R. (2004). Transforming Unstructured Content into "Meaningful" XML. Retrieved 8 January, 2004, from <http://www.dmreview.com/whitepaper/WID413.pdf>

Table 1. Comparison of data, information and knowledge

Supporting Literatures	Data	Information	Knowledge
(Quigley & Debons, 1999)	- symbols, numbers, or characters	- process, informed mental state, commodity, product, or thing	- as a thing
(Davenport & Prusak, 2000)	- set of discrete, objective facts about events - structured records of transactions in organizational context	- document or audible or visible communication - has meaning the "relevance and purpose" - data becomes information when its creator adds meaning by adding value	- a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information
(Spiegler, 2000; , 2003)	- record, store and maintain attributes	- when add value in some way	- when it adds insight, abstractive values, and better understanding
(Kaipa, 2000)	- symbols represent objects, events and/or their properties - out of context - no value until processed into useful forms	- a function of processed or structured data containing both the data and its relationship - provide objective descriptions - content oriented	- has both collective and personal components - has tacit and explicit nature - is the process of making sense of information

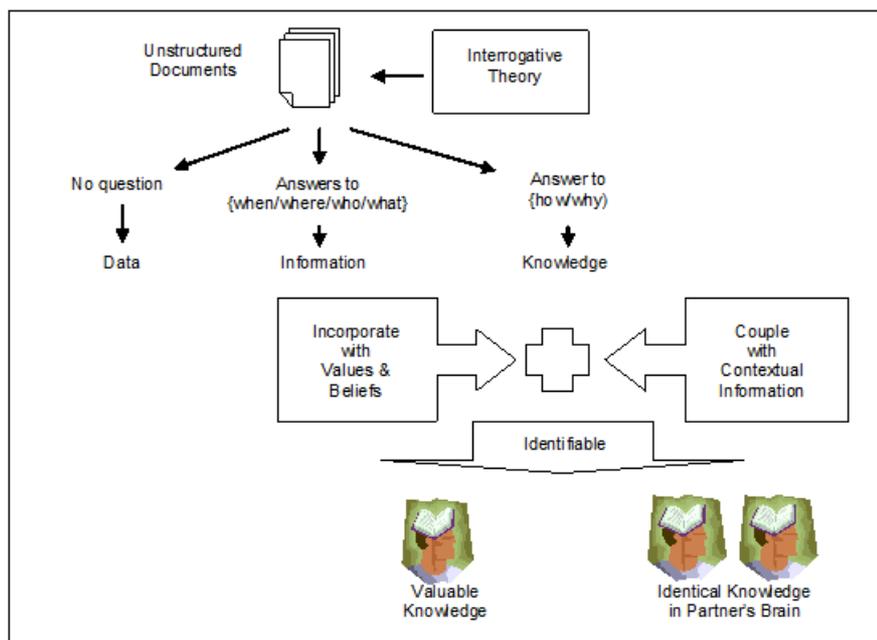


Figure 1. Interrogative Knowledge Identification Framework

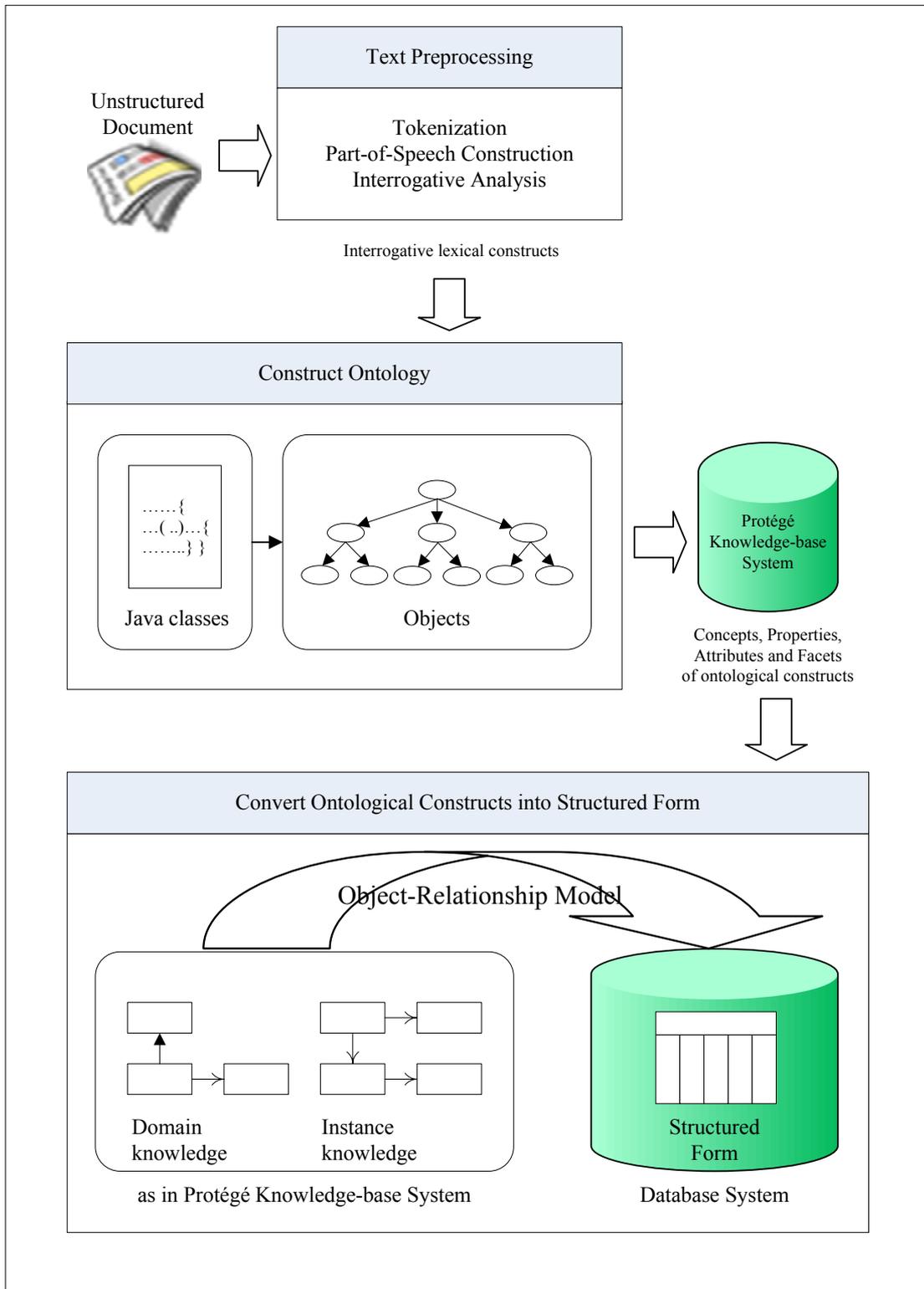


Figure 2. The MalayIK-Ontology Model