Intelligence Analysis Modeling

Ahmet Koltuksuz
Izmir Institute of Technology
ahmetkoltuksuz@iyte.edu.tr

Selma Tekir
Izmir Institute of Technology
selmatekir@iyte.edu.tr

Abstract

Intelligence is the process of supporting the policymakers in making their decisions by providing them with the specific information they need. Intelligence analysis is the effort of extracting the nature of intelligence issue with the policy goal in mind. It is performed by intelligence analysts who form judgments that add value to the collected material. With the increased open source collection capabilities, there has emerged a need for a model of intelligence analysis that covers the basic elements of valuable information: relevancy, accuracy, and timeliness. There exist models such as vector space model of information retrieval which only addresses the relevancy aspect of information and cannot cope with nonlinear document spaces. In this paper, we discuss the requirements of an integrated model of intelligence analysis along with its peculiar characteristics.

1. Introduction

The definition of power according to the Webster's dictionary is possession of control, authority and/or influence over others [17] although one may encounter some other definitions for different contexts such as in political sciences as the ability of one person to cause another to do what the first wishes, by whatever means [14].

One curious aspect of power is its constantly changing face. The land was power during the pre-industrial revolution times circa 1800s. And during the industrial revolution it revealed itself as being the steam-power as well as the electricity in early nineteenth century. The steel positioned itself as the new source of power during the world wars in between 1914 and 1945. Interestingly enough the nuclear power actually replaced the power of steel and thus marked the end of the second world war only to pave the way to cold war years in between 1950 and 1990 which were underlined by a nuclear power equilibrium. And it is the information this time as the most recent symbol of a power since 1980s which has been named as the information age. So, the newest format of the power is information.

And yet still a philosophical approach is also possible to the ever changing format of power as to whether the power is independent of our definitions for it or our definitions declare what the power is [3].

But from a different perspective the information itself is the collected and processed data for a specific purpose. And yet the intelligence is the tailored information in order to meet the demands of a specific customer. Or to put it boldly, the Intelligence is the process of supporting the policymakers in making their decisions by providing them with the specific information they need.

Stemming from the above definition is the intelligence analysis which represents the problem we try to address by providing some models created over the years. In this regard, the rest of this paper is organized as follows: Section 2 deals with the intelligence cycle. Section 3 compares and contrasts the intelligence analysis with the information retrieval. While the Section 4 briefly covers the information retrieval models, Section 5 specifically addresses the problem. Intelligence analysis model requirements covered in Section 6 and the conclusion is in Section 7.

2. The Intelligence Cycle

The intelligence process is stimulated by the question or request of the policymaker. Most of the time, the information repository in the agency is not sufficient to respond the policymakers' requests so there is a need to collect more information via the various collection disciplines. The collection disciplines are the means by which one gets access to

2006 International Conference on Hybrid Information Technology (ICHIT'06)
0-7695-2674-8/06 $20.00 © 2006 IEEE

information. The basic groups of collection disciplines are as follows:

- Human Based Intelligence (HUMINT)
- Imagery Intelligence (IMINT)
- Open Sources Intelligence (OSINT)
- Signals Intelligence(SIGINT)
    - Communications Intelligence (COMINT)
    - Electronics Intelligence (ELINT)
    - Radar Intelligence (RADINT)
    - Telemetry Intelligence(TELINT)

The technical collection disciplines (SIGINT) capture the information in transmission, are leaded by technological developments, and require special training and expertise for the personnel [11]. Human Intelligence (HUMINT) collectors are people carrying out covert action. HUMINT is needed as a major means of getting access to plans and intentions. Open source intelligence (OSINT) aims at producing intelligence using publicly available, low-cost information sources [15].

The collection phase is followed by the analysis phase. Intelligence analysis is the effort of extracting the nature of intelligence issue with the policy goal in mind. It is performed by intelligence analysts who form judgments that add value to the collected material. After the analysis phase, the intelligence product is presented to the intelligence customer-the policymaker. The policymaker evaluates the submitted intelligence and with his feedback to the intelligence analyst the intelligence process continues hence the intelligence cycle.

The emerging concept of open source intelligence has influences on the collection tasking. With the low-cost open source collection capabilities, intelligence analysts can be set freer to task the necessary collection. The distinction between collection and analysis weakens. The introduction of OSINT makes the intelligence analyst an intelligence collector as well. He utilizes information retrieval methods to gather information of interest plus describes his profile for information filtering purposes and constructs his own knowledge base in his working environment for future information retrieval.

3. Intelligence Analysis and Information Retrieval

Information retrieval (IR) is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data [16].

An IR system is intended to identify which documents the user should read in order to satisfy his information needs. For fulfilling this requirement, the system should implement three complementary tasks. In the first place, some method for the representation of what the documents are about (i.e., knowledge representation of documents) is needed. Secondly, closely related to the problem of representing the contents of documents there is the problem of characterizing the user needs. An additional problem in this context is one of making an assessment as to whether or not a document meets the actual needs of the user. That is, not only does the system have to represent the documents and user needs, but it must also provide a characterization of the process by which the user comes to a particular decision concerning relevance [19].

Intelligence analysis and information retrieval are two distinct concepts. In the first place, intelligence analysis is driven by the question of the policymaker in which there is a formal, explicit question to provide an outcome for. Intelligence analysts do the task of producing the outcome for the intelligence problem at hand. Secondly, information retrieval systems focus on the objective concerns but the person looking for information is of secondary importance.

Intelligence analysis systems, on the other hand, try to ease the job of the intelligence analyst by giving access to customizable and automatable analytical and collaboration tools including machine translation, knowledge discovery, trend analysis, and social-network analysis tools [8]. In this regard, the intelligence analyst plays a central role and, that is of the primary importance in intelligence analysis.

Exploration-discovery phase of the OSINT is of critical importance in that the nature of the intelligence problem is discovered at this phase. The components of analysis-insights are developed through the exploratory search process. In the context of intelligence analysis, distinctively different from IR which is based on an individual query, exploratory search can be characterized by successive queries interspersed with stages of sense-making. The search process itself not just the search results give the necessary insight. Thus, intelligence analysis systems should support capturing and visually representing analysts' iterative query processes and insights help them collect and compare information more effectively, as well as record and share the products of their analytic insights [5].

4. A Brief Summary of Information Retrieval Models

4.1 Entropy Based Models

Ensuring the accuracy of intelligence assessments is made difficult by the pervasiveness of uncertainty in intelligence information and the demand to fuse information from multiple sources. The uncertainty attribute of information is analyzed within the scope of the element of accuracy. Since Shannon, it's a widely accepted idea that uncertainty should be dealt with the use of the probability theory hence the entropy.

Generalized theory of uncertainty (GTU) proposed by Zadeh [21] enhances this idea by defining information as a generalized constraint on the values which a variable is allowed to take and probability theory fits within its conceptual structure as well. In GTU, uncertainty is linked to information through the concept of granular structure. In short, it recommends fuzzy set theory for representing and processing uncertain information.

Yager [20] introduces the measures of possibility and certainty as a tool to enable an intelligence analyst to provide an answer in terms of an upper and lower bound on the truth of the hypothesis and discusses methods for fusing information from multiple sources. The measure of the quality of the fused information is obtained by the "and" operation of the criterion of informativeness that is based on the measure of specificity by the criterion of credibility of the information source

The Information Fusiun and Uncertainty Analysis (Infusiun) software is a model-based tool that assists the intelligence analyst in aggregating information from multiple sources, identifying critical sources of uncertainty, and determining the impact of the uncertainty on the analyst's ability to assess potential threats in the environment [3]. The work makes use of fuzzy aggregation operations and the propagation of heuristic knowledge in belief models to do calculations within the model.

4.2 Enter Vector Space

4.2.1 Vector Space Model (VSM). Intelligence analysts do analysis on the collected material and thus support policymakers in their decision-making process. As part of his job, the intelligence analyst should provide the justification for the outcome he produced. He should base the outcome on the facts that are the results of his implication on the collected material.

There is a need to structure analysis because human mind is not capable of processing huge amounts of information and humans have mental traits that tend to lead them astray. Therefore, analytical tools will be helpful in structuring one's analysis. The net effect of which should result in separating the constituent elements of a problem in an organized way [7].

Restricted capacity of human’s information processing forces the reduction of the amount of information presented to the human. This situation gives rise to the need of efficient filtering and retrieval of information. The main barrier in this regard is the fact that machines are not able to understand human language. A number of technologies are available to structure/organize information in machine-understandable form. The main adopted approach is providing computable representations of natural language documents which have been explicitly developed for the solution of IF and IR problems. These approaches consider a small number of features of natural languages [10].

A number of technologies are available to structure/organize the retrieved information. The fundamental information retrieval approach organizes information into a vector space model (VSM) to represent each unique word within a document collection as a dimension in space, while each document represents a vector within that multidimensional space. Vectors that are close together in this multidimensional space form document clusters. The value of each vector coordinate is an entropy-based function of "local" and "global" frequencies of the word corresponding to that dimension. Local term frequency is the frequency of occurrence of terms within a document whereas global term frequency is the frequency over the entire set of documents. These term frequency counts are then used to calculate a weight for each term in each document, which is called the document term weighting. A pair wise comparison of each document in the system is made by the measurement of similarity between document vectors. The dot product (i.e., cosine of the angle between the vector pair) or Euclidean distance is used as the measure of similarity between two documents vectors [13].

Vector space model lacks in that it treats each index term as a separate coordinate and assumes the terms as being orthogonal, which is deemed contrary to the reality where term relationships exist and index terms are not assigned independently of each other [18].

Under the light of the fact that the index terms are not orthogonal but are related with each other, the problem of dimensionality should be investigated and a (vector) space of fewer dimensions than the number of distinct index terms should be identified. In other words, the intrinsic dimensionality of the document space might be much lower than the number of terms in the documents. Thus, many dimensionality reduction techniques have been applied to document representation and indexing.

4.2.2 Generalized Vector Space Model (GVSM). GVSM deals with the non-orthogonality of index terms by providing a method of computing the degree of similarity (or correlation) between non-orthogonal terms and the incorporation of this information into the retrieval strategy. It makes use of Boolean algebra to relate the non-orthogonal terms. The elements of Boolean algebra and Boolean expressions that represent terms are modeled as vectors in vector spaces [19].

4.2.3 Topic Based Vector Space Model (TVSM). TVSM is based on the dimensionality of the fundamental topics that is the orthogonal aspects in this case are the topics. Every term in this model is defined in terms of its relationship with one or more topics.

Every term is associated with a term vector and a term-weight. The term vector represents the position of term across the fundamental topics in the space of vectors and term-weight defines the closeness of one particular term to a particular topic [2].

5. The Problem

The information theory deals with the concept of information and in this regard classifies the information in three distinct types such as:

- Syntactic Information that is related to the characters from which the messages are built up.
- Semantic Information related to the meaning of messages and to their referential aspects,
- Pragmatic Information which is related to the usage of the messages [12].

The syntactic information is the only type for which we have an explicit mathematical definition with a unit. So it is our one and only yardstick as a measure for information. Such definition is known as the entropy which is a measure of uncertainty of a random variable.

Definition. The entropy H(X) of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x) \log p(x) ,$$

where x is a discrete random variable and p is the probability assigned to it and, since the log is to the base 2, the entropy is expressed in bits [12].

But when it comes to define a metric for semantic and/or pragmatic information which are the essential types for intelligence analysis it becomes another solid issue for fact that there is neither a mathematically proven theory nor any definition for them. This furthermore means that we do not know what we are trying to measure. So now, some questions can be put forward such as [9]:

- What exactly is semantic information?
- How can we measure it? In what unit?
- Is it continuous or discrete?
- Is there any proof that it is discrete? Or continuous?
- Is it deterministic or stochastic?
- Would it be possible to process it in a finite state machine if it is continuous and/or stochastic?
- How many dimensions will be needed for the definitions if it is continuous?

Moreover; the value of information when it is conceived in a semantic/pragmatic way creates a new issue as there are three basic elements which-taken together-comprise it: Information context (relevance), accuracy, and timeliness. Relevance is a subjective notion, and information retrieval systems do not deal with the subject at all. It is contradictory to seek relevant information which is subjective in nature without paying enough attention to the person who evaluates the context of information. Conventional information retrieval (IR) is limited in their help because of their disregard for personalization [1]. Information retrieval systems address only the relevancy aspect of information through the use of index terms-keywords. The more the count of a keyword in a specific document the more relevant the document becomes. However, this assumption is weak in that traditional query systems that are keyword driven are poor in representing the user intentions, requests.

6. Model Requirements

Most of the methods in information retrieval assume that the document space has a linear structure. However, there is no convincing evidence whether the document space is Euclidean or not. Therefore, it is more natural and reasonable to assume that the document space is a manifold, either linear or nonlinear [6]. It is necessary to devise dimensionality reduction techniques to estimate the intrinsic dimensionality of the document space where the document space is nonlinear.

Determining the granularity of information is the basic starting point for proposing an approach to structure the information. In the vector space model information retrieval objects are terms, documents, and queries. In the Infusion, the basic elements are a set of facts, a set of excerpts and a set of sources. In GTU, it's possible to define user defined granularities on the constrained variable.

The model should cover the basic elements of useful information which are relevance, accuracy and timeliness. Careful consideration reveals that the VSM of information retrieval deals with the relevancy aspect of information whereas the Infusion model covers the element of accuracy.

In the information retrieval systems, the success of the search process is dependent on the successful formation of the search query. The basic assumption behind this idea is that the person doing the search is aware of what he needs to know. The problem is that the famous "need to know" dictum only works in a world where you know what you need to know. In a world characterized by a high degree of flux and uncertainty, it is hard to know what a given analyst or consumer needs to know [8]. Thus, the model of intelligence analysis should permit local organization and storage of information under the direction of the intelligence analyst and should present innovative ways of search within the local storage.

In the analysis phase of the intelligence process, intelligence analyst first experiences the struggle of distillation-extraction of facts from the collected material and then production-formation of the alternative outcomes that are based on the extracted facts. The model should depict visually (if possible) the relationships between documents, excerpts and facts. Additionally, it should present the justification paths (fact-outcome relationships) for the produced outcomes. The proposed model should also be able to represent the documents in a nonlinear structure as well.

Relevancy of a document depends on many variables concerning the document as well as numerous user characteristics. Timeliness is one of the many factors that influence the judgment concerning relevance. Thus, vector-based approaches that target the relevance aspect at the same time should define time as part of the model. Document-term matrix representation of the knowledge base does not seem satisfactory as there is nothing to do with time. From this linear representation, it can be possible to define non-linear subspaces. However, every non-linear subspace that is estimated from the linear matrix representation cannot hold more information from the original linear representation. The key point here is that a non-linear representation (that takes into consideration time as well) of the natural language documents is required. In short, the original model should hold all the necessary information at the beginning.

7. Conclusions

So far all attempts to come with a definition for semantic and/or pragmatic information and even to define the information itself have been done in two dimensional Euclidean space without any result. It seems that trying to define the information in a two dimensional space as a scalar entity is not fruitful. The option left is to redefine information in a higher-dimensional space by Riemannian tensors [9]. Moreover, the analysis model should consider the time as yet another component and must find ways of integrating it to the aforementioned elements. In this regard, a manifold approach with Riemannian curved space seems much more relevant.

Once this is overcomed we will be in the need of new methodologies and software tools as well as possible new policies for the whole intelligence cycle.

8. References

[1]. Alonso R., Li H., "Model-guided information discovery for intelligence analysis", *Proceedings of the 14th ACM international conference on Information and knowledge management.*, Bremen, Germany, 2005, pp.269-270.

[2]. Becker, J., Kuropka, D., "Topic-based Vector Space Model", *Proceedings of the 6th International Conference on Business Information Systems*, Colorado Springs, 2003, pp. 7-12.

[3]. Boal, A., *Theater of the Oppressed*, New. Ed., Pluto Press, 2000.

[4]. Chopra K., Haimson C., "Information Fusion for Intelligence Analysis", *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, pp.111.

[5]. Gersh, J., Lewis, B., Montemayor, J., Piatko, C., Turner, R., "Supporting Insight-based Information Exploration in Intelligence Analysis", *Communications of ACM. 49, 4,* 2006, pp.63-68.

[6]. He X., Cai D., Liu H., Ma, W., "Locality preserving indexing for document representation", *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 2004, pp.96-103.

[7]. Jones M.D., *The Thinker's Toolkit: 14 Powerful Techniques for Problem Solving,* Three Rivers Press, 1998.

[8]. Kamarck E.C., "Transforming the Intelligence Community: Improving the Collection and Management of Information", *Transformation of Organizations Series.* IBM Center for The Business of Government, 2005.

[9]. Koltuksuz, Ahmet, "On Defining Security Metrics for Information Systems", *International Conference on*

Computational Methods in Science and Engineering - ICCMSE2005, *Brill Academic Publishers, Lecture Series on Computer and Computational Sciences, Vol. 4B,* 2005*,* pp. 1706-1707.

[10]. Kuropka, D, "A proposal for transformation of topic-maps into similarities of topics", http://kuropka.net/files/Topic-Sims.pdf.

[11]. Lowenthal M.M., *Intelligence from Secrets to Policy*, CQ Press, 2000.

[12]. Lubbe, Jan, *Information Theory*, Cambridge University Press, 1997.

[13]. Potok T.E., Elmore M., Reed J., Sheldon F.T., "VIPAR: Advanced Information Agents Discovering Knowledge in an Open and Changing Environment", *IIIS Agent Based Computing*, Orlando, 2003.

[14]. Shively, W. P., *Power & Choice, An Introduction To Political Science*, 4th. Ed., McGraw-Hill Inc., 1995.

[15]. Steele R.D., "Open Source Intelligence: Professional Handbook 1.0.", *Proceedings of the Fifth International Symposium on Global Security & Global Competitiveness Open Source Solutions*, 1996.

[16]. http://www.wikipedia.org, as of June 2006.

[17]. Webster's New Collegiate Dictionary, G. & C. Merriam Co., 1979.

[18]. Wong S.K.M., Raghavan V.V., "Vector Space Model of Information Retrieval-A Reevaluation", *Proceedings of the 7th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Cambridge, England, 1984, pp.167-185.

[19]. Wong S.K.M., Ziarko W., Raghavan V.V., Wong P.C.N., "On modeling of information retrieval concepts in vector spaces", *ACM Trans. Database Syst. 12, 2*, 1987, pp.299-321.

[20]. Yager R.R., "Fuzzy Set Methods for Uncertainty Management in Intelligence Analysis: Research Articles", *Int. J. Intell. Syst. 21, 5*, 2006, pp.523-544.

[21]. Zadeh L.A., "Toward a generalized theory of uncertainty (GTU): an outline", *Inf. Sci. Inf. Comput. Sci. 172, 1-2 Elsevier Science Inc.*, 2005, pp.1-40.