ARTICLE

TEXTS AND DATA MINING AND THEIR POSSIBILITIES APPLIED TO THE PROCESS OF NEWS PRODUCTION

WALTER TEIXEIRA LIMA JUNIOR
Casper Libero, Brazil

ABSTRACT The proposal of this essay is to discuss the challenges of representing in a formalist computational process the knowledge which the journalist uses to articulate news values for the purpose of selecting and imposing hierarchy on news. It discusses how to make bridges to emulate this knowledge obtained in an empirical form with the bases of computational science, in the area of storage, recovery and linked to data in a database, which must show the way human brains treat information obtained through their sensorial system. Systemizing and automating part of the journalistic process in a database contributes to eliminating distortions, faults and to applying, in an efficient manner, techniques for Data Mining and/or Texts which, by definition, permit the discovery of non-trivial relations.

Key-Words: Journalism; data mining; knowledge; technology, cognitive sciences

INTRODUTION

Why are human beings interested in news or journalistic information? That is an important question to be resolved in the field of journalism, and may be far from having a scientific answer. It can be noticed that in all countries, regardless of cultural differences, of democratic or authoritarian political systems, and of educational levels, among other comparative aspects, there is at least one mediatic system based on radio diffusion or printed media. This human necessity for knowing what is happening around oneself, and currently, with the advent of telematic networks, anywhere in the world, is the force which legitimates the profession called "journalist", making it necessary for society.

> Journalism is informing current facts, duly interpreted and transmitted periodically to society, with the objective of disseminating

> knowledge and orientating public opinion in terms of common well-being promotion. (BELTRÃO, 1992)[1]

Nonetheless, it is common knowledge that human beings are not only interested in news. Information absorption is a basic and survival need, and news in its industrial formatting for mass consumption is a category of the "information" class. Journalistic information formatted for different means (text, images and audio) is impregnated with the concepts of usefulness and veracity, and is absorbed by humans beings through their sensorial receptors (sight, touch, smell and hearing). Only taste is not employed in journalistic editorial products.

That kind of information mentioned above is important for humans, since it is used as an uncertainty reducer by the brain. Therefore, it composes, with other types, rules to be used in a moment of decision. The cognition process of that information travels along no other path than that of its absorption by memory.

In order to receive a wide range of loads and types of information, human beings are endowed with a system constituted of sensorial apparatuses. According to modern science, there is nothing in our minds from the surrounding world which has not gone through our sensorial apparatus.

The sensorial system is part of the brain system, which is responsible for processing sensorial information. The sensorial system encompasses sensorial receptors, neural links, and parts of the brain involved in sensorial perception. Sight, hearing, somatic sensation (touch), taste and smell are commonly recognized as sensorial systems, but there are others, like balance perception.

It is that system which allows us to have a particular representation of reality and performs the translation of all the information captured in the environment. Nevertheless, nowadays, in addition to collecting information from the environment about the fact, the journalist can collect data through the means of communication, use network collecting technologies (Internet) and traditional ones, like the telephone.

Thus journalism is a practice in which a human being (journalist) gathers information from the environment and/or through collecting technologies, and following technical and market criteria, tries to translate this representation of the real through communication platforms in its respective languages (printed, electronic, and digital) to other human beings (readers, radio listeners, viewers, and Internet users) who use their sensorial system to absorb that information, which is memorized or disregarded according to interests, cultural profiles and personal backgrounds. In those two stages of absorption, the informative reduction imposed by the perceptual chain, from journalist to final user of the information, becomes evident.

In Psychology and Cognitive Sciences, perception is the process of acquisition, interpretation, selection and organization of sensorial information. The word "perception" comes from Latin: *capere*, which means "to take", and the prefix *per* which means "completely".

The methods of studying perception vary from essential approaches to Biology or Psychology, to psychological approaches to the philosophy of the mind and Empirical Epistemology, such as the ones proposed by David Hume, George Berkeley or Merleau Ponty's assertion, considering perception as the basis of all sciences and knowledge.

Is it possible to emulate?

In the field of applied Social Sciences there are many questionings which come from researchers in the area about whether it is possible for computational machines to receive data and run programs that select, find or hierarchize, even in a reduced way (modeling), and by approximation to the human mental process, journalistic news within the "news values" pinpointed by researchers of the area, such as Gislene Silva and Érica Frazon.

> When facts in the material production of a piece of news are treated journalistically, selection and imposition of hierarchy resort to news values. But they act here just as a part of the process, since in these sequenced choices other criteria for a fact to become news are involved, such as product format, image quality, editorial line, cost, public target, etc. News values, the characteristics of the fact in itself, in its origin, are only a subgroup of facts acting together with this second group of news potentiality criteria, related now to the treatment of the fact. Studying selection also implies tracing the judgments of each selector, the organizational, social and cultural influences they suffer when making their choices, the various agents of those choices who occupy different positions on the editorial staff, and even the participation of sources and of the public in those decisions – it is worth mentioning here the studies of agenda-setting, which complexify the investigations about the process of news selection. (SILVA, 2005, p. 5)

In spite of all the complexity of the issue pointed out by the researcher from the Federal University of Santa Catarina, her systematization of news potentiality criteria partially reveals how news values are constituted. It is a reverse engineering task of human thinking, discovering to a slight extent what makes a human being pay more or less attention to a piece of news (see appendix). In other words, the task is an initial attempt to structure and classify attributes and their respective scales of news attributes. This type of organization may work as a basis for initiating model simulations, using computational systems with the help of databases. This assertion is based on the premise that there is a human logic in the search for news. Logic is not the privilege of some area of knowledge, it permeates all human activity. Logic, for example, has turned into the basic language of formal sciences. Computational systems and software are products of formal sciences.

> A science is any organized body of knowledge which has principles. The first principles of any science are those fundamental truths on which they rely and on which all their activities are based. Logic, as a science, has its fundamental principles, but Logic keeps a unique relation with all the other sciences, since the first principles of Logic apply not only to Logic, but to all sciences. As a matter of fact, its bases are wider, because they are applicable to human reason as such, although it must be exercised (MCLNERNY, 2006, p. 46)

Identifying parameters

The empirical process and years of refinement have established a connection between what the user understands as news and what is transmitted by journalists through the media. That connection is established by the "news attributes" mentioned above. Therefore, building a Knowledge Basis (KB) interconnecting previously diverse areas is not an easy task.

First and foremost, it is necessary to articulate people (GARCIA, FLÁVIO, FERRAZ, 2005). Establishing synergy between knowledge engineers and specialists in a domain requires dedication, planning and a dose of goodwill.

While knowledge engineers are responsible for transforming ideas, concepts and ways of rationalizing the world into a model processed by computers, the specialist in a domain needs to strive to translate his knowledge into clear and objective language, besides evaluating and pointing out the mistakes made by the system, based on the answers obtained. In other words, it is a continuous cycle of improvement.

Nevertheless, in addition to the problem of synergy, there is another drawback impairing the success of a KB development, which is knowledge collecting. A myriad of techniques are employed to solve this problem, but none of them is perfect and, due to that, creativity is required from the knowledge engineer and patience from the specialist to correct the flaws. The commonest techniques are:

> Manual knowledge acquisition techniques based on interviews, accompaniments or models; semi-automatic acquisition techniques based on cognitive theories or on existing models; machine learning technology attempting to induce rules from catalogued examples; data mining technology which tries to extract rules and behavior from the analysis of large amounts of data; data mining technology which tries to extract knowledge from large amounts of non-structured data. (GARCIA, VAREJÃO, FERRAZ, 2005, P. 68)

However, after extracting and representing computationally the knowledge of a domain, work does not become easier. Some requirements need to be met for the Knowledge Basis to be efficient:

> It must: be comprehensible to the human being, for in case it is necessary to evaluate the knowledge state of the system, the Knowledge Representation must permit its interpretation; it must disregard the details of how the knowledge processor, which will make the interpretation, works internally; it must be robust, that is, it must permit its use even if it does not apply to all possible situations; it must be generalizable, contrary to knowledge itself, which is individual. A representation needs various points of view of the same knowledge, so that it may be attributed to diverse situations and interpretations. (REZENDE, PUGLIESI, VAREJÃO, 2005, p. 29)

More than systematizing and automating part of the journalistic process, the construction of a Knowledge Basis (KB) with the best practices allows one to compare records from the database with the established rules and to provide subsequent storage of the patterns encountered, which benefits other processes. Among them, we highlight the benefits of applying Text and/or Data Mining to help in the investigation, in the complementation and even in the scoop.

It is important to highlight that although Data Mining (DM) and Text Mining (TM) are widely discussed in the field of Computer Sciences, the effort to relate them to applications in journalism is recent. Therefore, compatibilization difficulties are naturally found.

Data Mining is concerned basically with searching for occult patterns in masses of data found in corporate data warehouses or in Knowledge Bases of Intelligent Systems. As DM is a concept which involves Statistics, Artificial Intelligence and Machine Learning, it pans for information of strategic value, which is "invisible" in the records, allowing the identification of tendencies for an anticipated view of future scenarios and the discovery of new data patterns, not always perceived by human analysts.

There are many definitions for Data Mining, but the one which has been most accepted is that of Usama Fayyad (1996).

> Database knowledge extraction is the non-trivial process of identification of valid, new, potentially useful and comprehensible patterns embedded in the data (FAYYAD, PIATESKY, SHAPIRO, SMYTH, 1996).

Just for comparison, Rowen & Cilione (2000) understand that "data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The process consists of four basic stages: (1) data preparation, (2) exploration, (3) model building (or pattern definition), and (4) validation/verification[3]".

In the essence of the two definitions, we realize that their usefulness is related to the development of automated systems for journalistic production, but as Walter Lima Jr (2006, p.125) explains, databases need to be precise, non-historical, and have a certain artificial intelligence to deal with the semantic modifications of words, for instance. Data mining makes it possible to extract valid patterns[4]; for example, to investigate whether unemployment rates decrease when elections approach and why that happens.

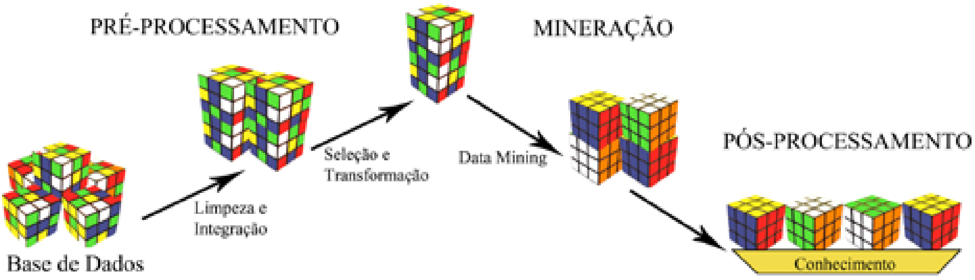
The possibility described above is pertinent for a single reason: even in relational databases, when well-designed, there is an extraction of diverse information by using Structured Query Language (SQL)[5]; however, the process necessarily requires the formulation of questions to be solved. No matter how creative the analyst is, he will be able to formulate only some questions for the system to map the database, and eventually bring practical results. In other words, due to the amount of data involved, patterns and relevant behaviors are ignored, and in this respect, Data Mining demonstrates its advantages.

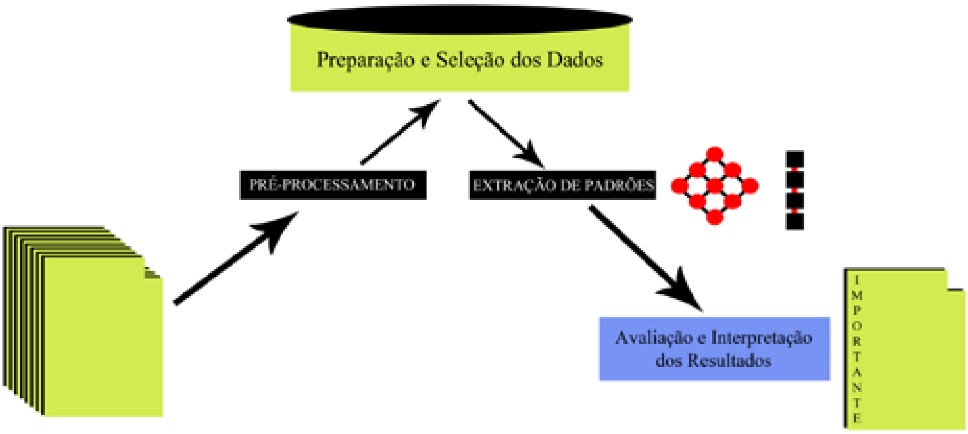
The mining process identifies, through *tasks* (classes of problems) and *techniques* (groups of solutions that use algorithms to solve the problems proposed in the tasks) the questions and answers in the database. To summarize, it is possible not only to relate events based on the record, but using that as a starting point, to act in a predictive way. It is important to highlight that we do not mean that the journalist will start giving credit to speculations or, worse, to invented facts with this kind of application, but the gain in precision on following the details of a development is remarkable.

Another relevant application to journalism is Text Mining (TM). According to Tan (1999), this new area is defined as a process for extracting interesting and non-trivial patterns or knowledge from a group of textual documents. The similarity with the definition of Usama Fayyad (1996) for DM is not a matter of chance, since the inspiration for TM actually came from the process of Data Mining. Nevertheless, unlike DM, which involves extracting information from structured databases, TM extracts information from non-structured or semi-structured databases. That difference allows us to work with diverse factors that cause a complexity of tasks, such as dealing with different types of language, style or content of the written document.

Overall, the two applications are discussed in various books and articles as part of a wider project called Knowledge Discovery in Databases (KDD), and Knowledge Discovery from Texts (KDT). Both are similar in terms of the treatment of data and involve many stages, like selection, pre-processing, transformation, data mining, report generating, as well as the interpretation of the results obtained. Some authors[6] present variations of these stages, but it is possible to group them as follows: pre-processing, mining and post-processing. Below, there are two graphs which illustrate the processes of KDD and of KDT:

[Following the arrows, the translation is as follows: collection of documents, pre-processing, preparation and selection of data, extraction of patterns, evaluation and interpretation of results]




Preparação e Seleção dos Dados
PRÉ-PROCESSAMENTO
EXTRAÇÃO DE PADRÕES
Avaliação e Interpretação dos Resultados
IMPORTANTE

[In the graphic process, there are three stages: pre-processing, mining and post-processing. Following the arrows, the translation is as follows: Database, Cleaning and Integration, Selection and Transformation, Data Mining, Knowledge]

In summary, pre-processing is the stage which aims at engendering a convenient representation for the mining algorithms, starting from the database. It includes the selection (automatic and/or manual of relevant attributes), sampling, transformations of representation, etc. Data mining is the application of algorithms to pre-processed data, and post-processing is the selection and ordering of the interesting discoveries, mappings of representation of knowledge, report generation and interpretation of users and/or specialists.

To understand the real advantage of using mechanisms which help in the process of investigation and complementation of material, we could think of two large Brazilian media groups that pioneered in the construction of databases: Grupo Abril and the group which includes the newspaper Folha de São Paulo. This is what the website of one of these groups says:

> ... Folha data bank is a journalistic collection which contains more than eight decades of recent Brazilian history. Its objective is to give support to journalists of the Folha da Manhã group and to give assistance to researchers, students and companies in their researches. The collection includes the newspapers edited by the group, an archive of extracts with about 100 thousand thematic folders and 20 million

images in physical and digital archives[7].

As for Grupo Abril, Walter Lima (2006, p.121) recalls that it has its Dedoc, inaugurated in 1968. “In the past, everything was manual. In 1984, a computerization process began. Veja magazine was the first to provide access to the summary of all the articles and the research of reference words. All the magazines of the group are in a database called Fólio News. At the time of the compilation, Veja magazine, the publisher´s flagship, for example, had 43,687 articles; Anamaria, 19,587; Exame, 12,958; and Cláudia, 11,262”.

In spite of so much information in the vehicles, there is no efficient and efficacious system for relating information (based on key words) which would permit journalists to use their capability of connecting information to carry out relatively simple tasks, such as tracing panoramas. Supposing that the deadline for preparing a special section about the history of a certain social movement since the 80’s is one day, and that the editor chooses only one journalist to undertake the activity, there will not be enough time.

The difficulty in this case is not lack of information, but how to make a selection in the best way possible and in a very short time. What is the best way to use the system in order for it to provide what is desired, without losing time in search of data that are not important to the context of the article? How can that which you desire be found with a single sentence? Or, why has the system not provided it automatically much earlier, regardless of the need for a journalistic subject in a short time?

What this essay suggests, not ignoring the editorial staff´s world, is that time is short and journalists need mechanisms to help them find relevant, precise and guiding information for their articles. Resorting to the “Archive Department” is no longer possible. Due to a constant policy of cost reduction, this kind of department has been eliminated. In the past, this was the best way to obtain a record of the main happenings. In the history of journalism, it is easy to find important articles which were written with the help of this kind of department.

Journalists and informative conections

Journalist Cláudio Júlio Tognolli, who describes himself as investigative, has used many services offered by the database of Editora Abril and, according to him, has also developed abilities to extract information from search engines (LIMA JR, 2006).

Walter Lima Jr (2006) affirms that Tognolli's technique is based on always starting to search in Google Images, not Google Text, since the former mechanism provides a chaotic subgroup of images which are more interesting than the other system:

> Therefore, if I have a specific report based on free associations about a person, and I want to search for this person on the Internet, I think about the person for some minutes and associate him/her with twenty or thirty words. It is quite simple. I type "his/her name", and add "and crime", "and car", "and guitar", based on my view of that person. So, I create a scheme by booleaning, using "and" with free associations (Tognolli, 2004)

The drawback in the aforementioned example is that each professional has a refined technique to reach his/her objective. With the use of applications such as Data Mining or Text Mining, that technique would be reinforced and would update some potential aspects which make digital journalism[8] possible, as the memorizing and personalization or customization of contents illustrate.

Text Mining, for instance, makes it possible to exhibit a structured representation of documents, frequently in the format of an attribute-value chart. This chart is characterized by its high dimensionality, as each term of the document may represent a possible element in the group of attributes of the chart. Therefore, data selection is crucial in order to reduce the extension of the attribute-value chart, so that truly significant data are obtained.

To attain that goal, some tasks are carried out during the pre-processing stage, such as a lexical analysis, an elimination of irrelevant terms or stopwords[9], as well as a morphological normalization of terms (removal of prefixes and suffixes).

But, after all, how can Text Mining help Journalism through a structured representation of documents? How can essential terms be found? We can anticipate that the answer is not simple. In general terms, there are two approaches in order to work with textual data: one of them is semantic analysis, the other is by means of statistics. The first involves an evaluation of the sequence of the terms in the context of the sentence, whereas the latter concerns the counting of the number of times a term appears in the text.

Semantic Analysis

Using fundamentals and techniques based on the processing of natural language, proper semantic analysis, when added to a more complex linguistic processing, identifies correctly the function of each term. For that to happen, some types of knowledge are necessary:

> **Morphological:** knowledge of the structure, form and inflection of words. **Syntactic:** structural knowledge of lists of words and of how words can be matched to produce sentences. **Semantic:** what words mean irrespective of context, and how more complex meanings are produced by the combination of words. **Pragmatic:** knowledge of language use in different contexts, and how meaning and interpretation are affected by the context. **Discursive:** how immediately preceding sentences affect the interpretation of the next sentence. **World:** general knowledge of the domain or the world to which the communication of the natural language is related. (EBECKEN, LOPES, COSTA, 2005, p. 339-340)

There are basically two problems with that approach: customizing the application according to the language and a high rate of errors when one tries to reach the essence of a figure of speech. Nevertheless, the second problem is alleviated by the justification that good journalistic practice avoids ironies, metonymies and all kinds of structures which may put good understanding at risk.

Statistical analysis

Statistical analysis entails knowledge derived from a quantitative study of the terms. The main advantage of this application is that the strategy can be used in any language. However, as repetition of terms is not advisable in a journalistic text, methods that overcome this obstacle must be used. The solution is to specify a dictionary or thesaurus as a controlled vocabulary which represents variable terms, such as synonyms, abbreviations, acronyms and spelling alternatives (EBECKEN, LOPES, COSTA, 2005, p. 349). It must be pointed out that this alternative can also be applied to a semantic approach, but the time needed for processing the statistical analysis is shorter with relation to large amounts of texts.

Both Text Mining and Data Mining resort to diverse classes of tasks to discover non-trivial relations and in this way efficiently achieve the proposed objective. In this specific case, we want applications to act in the investigation, in the complementation and even in the scoop.

Ebecken, Lopes & Costa (2005) highlight the following tasks of Text Mining:

> The process of **clustering** makes the relation among documents explicit, while **categorization** identifies the key topics of a document. **Extraction of characteristics** is used when it is necessary to know people, places, organizations and objects mentioned in the text. **Summarization** expands the principle of extraction of characteristics, concentrating more on whole sentences rather than nouns or phrases. **Thematic indexation** is useful when one wants to be able to work preferably with topics instead of key words. (EBECKEN, LOPES, COSTA, 2005, p. 351) [the words in boldface were highlighted by the author of this article]

> The task categories of Data Mining are divided into predictive and descriptive.
>
> **Predictive** – Classification: the task aims at gathering data in pre-defined classes. Estimation (regression): this aims at defining the value (numerical) of some unknown variable based on the values of known variables. **Descriptive** – Association: this studies a pattern of relation among items of data. It is used to identify patterns in historical data. Clustering (segmentation): information can be partitioned in classes of similar segments. In this case, no information is given the system regarding existing classes. The algorithm itself discovers the classes starting from alternatives found in the database, thus grouping a set of subjects in classes of similar subjects. (VIANA, 2004, p.18)

This type of treatment of data and the other types mentioned throughout this essay show that there is a real possibility for transmission of procedures and patterns, already used in other areas, for a group of digital systems of storage and of relation between data and information, aiming at the building of knowledge in the field of journalism.

Final considerations

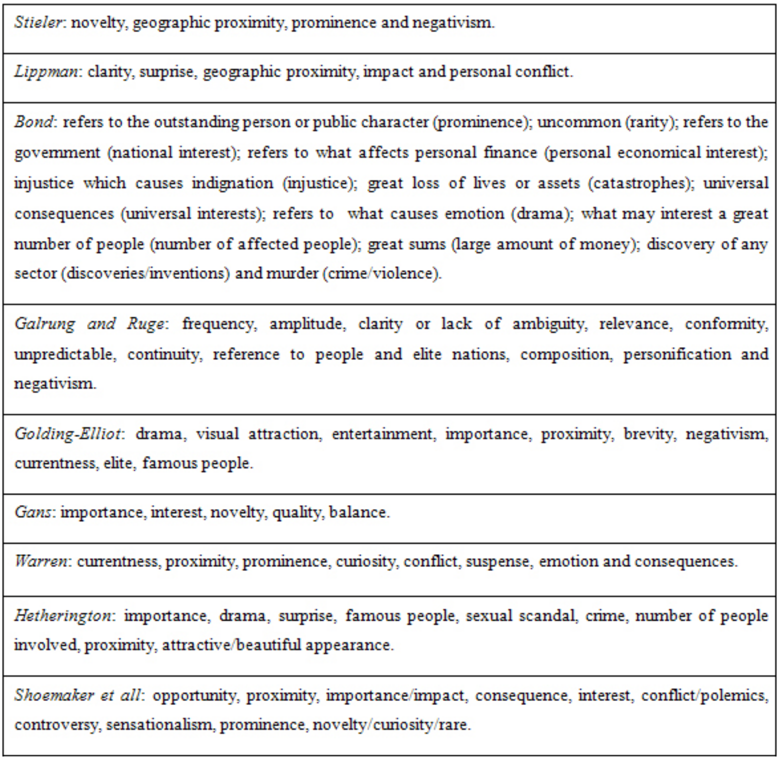
The use of technologies in the field of journalism is not something new and astonishing. Nevertheless, with the advent of digital supports through databases and software that treat information in a brisk, relational, customized and hierarchized way, journalistic practice gains new possibilities and expands its capability for a qualitative and informative treatment of news. The processes of investigation, complementation and search of relevant and new journalistic facts, for example, are revigorated due to an effort in the building of this Knowledge Basis resting on the best practices in the field, and on the application of up-to-date technological tools, such as Data Mining and Text Mining.

What we propose is a reverse engineering work of human thinking that reveals a little more what leads a human being to pay more or less attention to a piece of news.

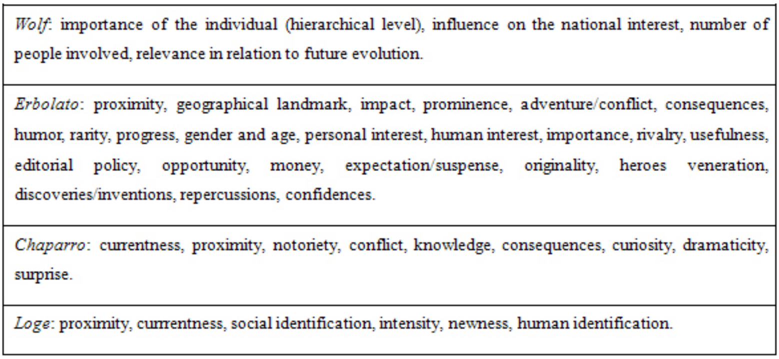
In other words, it is an initial attempt to structure and classify attributes and their respective scales of "news values". This type of organization may be used as a basis to initiate simulation of models using computational systems with the help of databases. This assertion presupposes that there is a human logic in the search of news.

In this sense, we believe that using consolidated knowledge in the fields of Computer Sciences and Cognitive Sciences to solve systemic problems of the rudimentary processes of the production of journalism in databases, for example, will enable journalism practices and objectives to be equated and harmonized with a globalized society which already has, in some parts, multimedia apparatuses to obtain information with high computational power, in real time, in high definition, wireless, on the way toward the knowledge society.

APPENDIX



Stieler: novelty, geographic proximity, prominence and negativism.
Lippman: clarity, surprise, geographic proximity, impact and personal conflict.
Bond: refers to the outstanding person or public character (prominence); uncommon (rarity); refers to the government (national interest); refers to what affects personal finance (personal economical interest); injustice which causes indignation (injustice); great loss of lives or assets (catastrophes); universal consequences (universal interests); refers to what causes emotion (drama); what may interest a great number of people (number of affected people); great sums (large amount of money); discovery of any sector (discoveries/inventions) and murder (crime/violence).
Galrung and Ruge: frequency, amplitude, clarity or lack of ambiguity, relevance, conformity, unpredictable, continuity, reference to people and elite nations, composition, personification and negativism.
Golding-Elliot: drama, visual attraction, entertainment, importance, proximity, brevity, negativism, currentness, elite, famous people.
Gans: importance, interest, novelty, quality, balance.
Warren: currentness, proximity, prominence, curiosity, conflict, suspense, emotion and consequences.
Hetherington: importance, drama, surprise, famous people, sexual scandal, crime, number of people involved, proximity, attractive/beautiful appearance.
Shoemaker et all: opportunity, proximity, importance/impact, consequence, interest, conflict/polemics, controversy, sensationalism, prominence, novelty/curiosity/rare.



Wolf: importance of the individual (hierarchical level), influence on the national interest, number of people involved, relevance in relation to future evolution.
Erbolato: proximity, geographical landmark, impact, prominence, adventure/conflict, consequences, humor, rarity, progress, gender and age, personal interest, human interest, importance, rivalry, usefulness, editorial policy, opportunity, money, expectation/suspense, originality, heroes veneration, discoveries/inventions, repercussions, confidences.
Chaparro: currentness, proximity, notoriety, conflict, knowledge, consequences, curiosity, dramaticity, surprise.
Loge: proximity, currentness, social identification, intensity, newness, human identification.

Source: SILVA, Gislene. Valores-notícia: atributos do acontecimento. Trabalho apresentado ao NP 02 – Jornalismo, do V Encontro dos Núcleos de Pesquisa da Intercom, 2005

NOTES

1 All the quotations in this paper were freely translated from the original article written in Portuguese.

2 Data warehouse is a computational system used to store information related to the activities of an organization in data banks, in a consolidated form. The design of the data basis favors reports and analyses of large amounts of data, and acquirement of strategic information which can make decision making easier. The data processing in a data warehouse is always referred to as Online Analytical Processing (OLAP), in contrast with Online Transaction Processing (OLTP) – used to store business operations. Another difference is that the data in a data warehouse are not volatile, that is, they do not change, unless it is necessary to make corrections of previously loaded data. The data in this case is only for reading, they cannot be altered.

The data warehouse allows the analysis of large amounts of data, stored by the transactional systems (OLTP). They are what is called historical series, which make it possible a better analysis of events for present decision making and the prediction of future events.

Due to its capacity to summarize large amounts of data and make analysis possible, data warehouses are currently the core of management information systems and a support to the decision of the main business

intelligence solutions of the market.
Available at: http://pt.wikipedia.org/wiki/Data_warehouse Accessed on June, 27th, 2007

3 (Rowe, Cilione, 2000) For more information: ROWE, Ken; CILIONE, Patrick. *Data Mining and Neural Networks Analysis*. Available at: http://web.archive.org/web/20050616182712/http://acspri.anu.edu.au/newsletter/news42/DataMining.htm Accessed on June, 27th, 2007

4 Patterns are repetitive information units, or sequences of information which have a repetitive structure.

5 Structured Query Language is declarative research language for relational data banks. Most of the original characteristics of SQL were inspired in relational algebra.

6 Such as: Rezende, Pugliesi, Melanda & de Paula (2005) e Fayyad (1996)

7 Available at: http://bd.folha.uol.com.br/ Accessed on July, 10th, 2007

8 Digital journalism is every discursive product which builds up reality through the uniqueness of events, which has as its support telematic nets or any other kind of technology through which numeric signs are transmitted, and which incorporates the interaction with users throughout the productive process. It is one of the activities that are developed on cyberspace (MACHADO, 2000, p.19). In relation to the characteristics which define digital journalism, Schwingel (2005j, p.01) says the following: it is composed of hypertextuality, multimediality, continuous updating, memory, personalization or customization of content, interactivity and suppression of space and time limits to the posting of information in its first nature.

9 Ebecken, Lopes & Costa (2005, p. 347) say that stopwords or stoplist is the name give to the removal of words or terms which do not contain knowledge in the text, such as , auxiliary and connective words (and, for, at, they) that do not translate the essence of texts.

10 Such as: Rezende, Pugliesi, Melanda & de Paula (2005) e Fayyad (1996)

BIBLIOGRAPHY

ALVARES, Reinaldo V. *Mineração de Dados: Introdução e Aplicações*. Artigo publicado na revista SQL Magazine, edição 10, ano 1, 2004

BELTRÃO, Luiz. *Iniciação à filosofia do jornalismo*. São Paulo: EDUSP, 1992

DA SILVA, Cassiana F. *Uso de Informações Linguísticas na etapa de pré-processamento em Mineração de Texto*. Dissertação de mestrado defendida no Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio do Sinos, São Leopoldo (RS), 2004.

DAVIS, R., SHROBE, H. and SZOLOVITS, P. *What is a Knowledge Representation?* AI Magazine, v.14, no.1, p. 17-33, Menlo Park, USA. 1993 Disponível em: http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html Acesso em 23 de junho de 2007

HAN, J., KAMBER, M. *Data mining: concepts and techniques*. USA: Morgan Kaufmann, 2001.

KORFHAGE, Robert R. *Information Retrieval and Storage*. New York: John Wiley & Sons, p. 349, 1997.

KOWALSKI, Gerald. *Information Retrieval Systems: Theory and Implementation*. Boston: Kluwer Academic Publishers, p. 282, 1997.

LAGE, Nilson. *Ideologia e técnica da notícia*. Florianópolis: Ufsc-Insular, 2001.

LEME, Maria Isabel da Silva. Aquisição de conhecimento. Bol. psicol. [online]. dic. 2005, vol.55, no.123, p.233-239. Disponível em: http://pepsic.bvs-psi.org.br/scielo.php?script=sci_arttext&pid=S0006-59432005000200008&lng=es&nrm=iso Acesso em 24 de junho de 2007

LIMA JR. Walter Teixeira. *Jornalismo Inteligente na era do* data mining. Publicado na Revista do Programa de Pós-graduação da Faculdade Cásper Líbero, ano IX – no.18, p. 121-126, 2006.

MACHADO, E. *La estructura de la noticia en las redes digitales: un estudio de las consecuencias de las metamorfosis tecnológicas en el periodismo*. Tese de doutorado defendida no Programa de Doutorado em Jornalismo e Ciências de Comunicação da Universidade Autônoma de Barcelona. Barcelona (Espanha), Junho de 2000.

MCLNERNY, D. Q. *Use a Lógica*. Rio de Janeiro: Best Seller, 2006.

REZENDE, Solange (org) *Sistemas Inteligentes – Fundamentos e Aplicações*. Barueri, São Paulo: Manole, 2005

SALTON, G,;MACGILL, M. *Introduction to Modern Information Retrieval*. New York: McGRAW-Hill, p.448, 1983.

SCHWINGEL, Carla. *Jornalismo digital de quarta geração a emergência* *de sistemas automatizados para o processo de produção industrial no Jornalismo Digital*. Apresentado no GT de Estudos de Jornalismo da Compós em Porto Alegre (RS), 2005.

SILVA, Gislene. *Valores-notícia: atributos do acontecimento (Para pensar critérios de noticiabilidade I)*. Trabalho apresentado ao NP 02 - Jornalismo, do IV Encontro dos Núcleos de Pesquisa da Intercom, Porto Alegre, 2004.

TAN, A. *Text mining: the state of the art and the challenges*. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD'99, p. 65-77, Beijing, april 1999.

TOGNOLLI, Cláudio Júlio. Entrevista concedida a Walter Teixeira Lima Júnior em 10 de setembro de 2004.

Walter Teixeira Lima Junior is a professor in the Master's Program of Casper Líbero College, with a postdoctoral degree in Communication and Technology (Methodist University) and Doctor in Communication Sciences (São Paulo University). E-mail: digital@walterlima.jor.br