From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art

Nabil Khemiri, Sahbi Sidhom

From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art

Nabil Khemiri, Sahbi Sidhom

From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art.

Nabil KHEMIRI
FSJEGJ
University of Jendouba
Tunisia
nabil.khemiri@fsjegj.rnu.tn

Sahbi SIDHOM
LORIA
University of Lorraine
France
Sahbi.sidhom@loria.fr

***Abstract*— In the era of Big Data and Open Data, a massive and heterogeneous collections of documents (from text to multimedia) are created, managed and stored electronically. *to make these documents more usable*, a manual and/or automatic indexing process allows to create a representation of documents by a set of metadata, descriptors and social tags. These representations then make it easier to find information in a massive and scalable collection of documents from different sources (social networks, open data, …) to respond to user information needs (user requests). Numerous research studies have been carried out to propose indexing approaches depending on the type of indexed documents. Also, the evolution of indexing Methods, documents representation, electronic content, Big Data and Open Data. This paper presents a state of the art of approaches and methodologies ranging from manual and automatic indexing to algorithmic methods in the era of Big Data and Open Data.**

***Keywords*—Big Data, Open Data, indexation (intellectuelle, automatique), recherche d'information, document (contenu hétérogène, multimédia). Big Data, Open Data, manual indexing, automatic indexing, information retrieval, heterogeneous content, multimedia document.**

I. Introduction

Nowadays, Information occupies a central place in our daily life. It represents a source of knowledge and power. In the era of Big Data and Open Data, a huge amount information, documents, multimedia content and social tags are created, managed and stored electronically. Which explains the exponential growth of data flows from a wide variety of fields that have led to the creation of an unprecedented amount of data. With this huge amount of data, it is becoming increasingly difficult to respond to user queries who looking for relevant documents results. This is why new methods and algorithms have emerged to better represent the information collected from heterogeneous sources. In order to make these documents usable, a manual and / or automatic indexing process allows to create a document representation by a list of metadata, descriptors and social tags. These representations are used to find relevant information in a scalable collection of documents, to response to user requests (information needs). In this context, numerous research works have been carried out to propose indexing approaches. The ultimate goal of these different approaches is to better represent contents (documents, electronic content, Big Data and Open Data) to effectively identify those that are most relevant when searching for information. This paper presents a state of the art of approaches and methodologies ranging from manual and automatic indexing to algorithmic methods in the era of Big Data and Open Data.

II. Indexing definition

Indexing is a process of representing information which consists in identifying the significant elements to characterize a multimedia documents (i.e. audio, images, text, video). This process analyzes documents to assign or extract a list of descriptors, metadata and social tags. These representations of will subsequently facilitate the search for information in a collection of documents. Knowing that an Information Search System (IRS) must be composed of three principal functions (cf. Figure 1):

(i) Represent documents content
(ii) Represent user information needs (query*) and*
(iii) matching process: compare these two representations in order to find documents, order the search results by relevance, and return the documents to user.

Consequently, the performance of the IRS depends on the choice of representation model and the matching process.

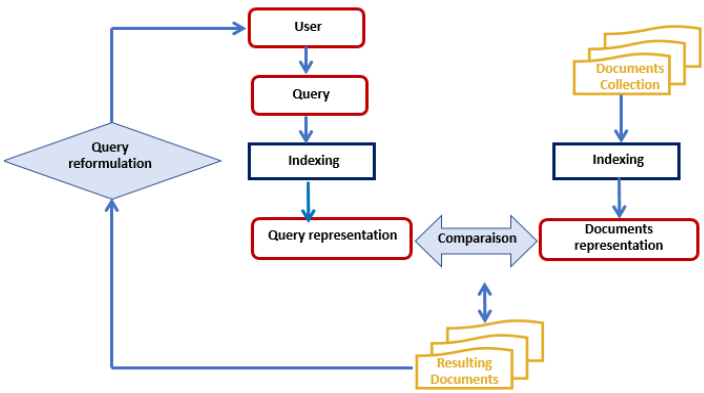

Figure 1: **Information Search System**

There are three approaches that can be distinguished: the manual indexing approach, the automatic indexing approach

and the semi-automatic indexing approach (combining an automatic approach with a manual approach) of documents.

III. The Manual indexing

Manual indexing (intellectual or human) is based on associations between words in document with controlled vocabulary terms (manually assigned indexing terms). The choice of terms that represent each document (descriptors) depends on the know-how of the indexer, his knowledge and practical experience in the field of indexing. The human indexer uses a documentary language such as the thesaurus which provides a hierarchical dictionary (controlled vocabulary) of pre-established monolingual or multilingual standard terminologies to index documents. This type of approach allows classification and research by concepts (subjects or themes) in a collection of documents. Manual indexing is the result of a document content analysis which is based on the following four steps:

- **Documentary analysis** (**document analysis**): the indexer must have a global knowledge of the document to be analyzed. To analyze a document, he first consults the title, the table of contents, the summary, the introduction, the introductions and the conclusions of the chapters (if they exist) and the conclusion. That speed-reading allows the indexer to know the main subject (theme) discussed or described in the document.
- **The choice of concepts (keywords):** to define the main concepts that best characterize a document, the indexer must answer a certain number of questions, those that a user would ask when searching for information such as: who and what the document is about? where and when?
- **Conversion of concepts into descriptors:** the indexer chooses the appropriate index terms (the descriptors) from a controlled vocabulary list. A controlled vocabulary is finite set of index terms from which all index terms must be selected. Only approved terms can be used by the indexer to describe the document. which ensures uniformity in the representation of the document [1].
- **Proofreading and revision:** during this step, the indexer decides to retain or reject some descriptors.

Human indexing has several disadvantages. It is too costly in terms of money and vocabulary building time and assignment of concepts (index terms) to documents. It is subjective, since the choice of indexing terms depends on the indexer and his level of knowledge of the target domain. although, indexers follow the same steps, different concepts can be selected to characterize the same document.

Also, controlled documentary language is difficult to maintain since the terminology is constantly evolving. When there is a *high* volume of documents, manual indexing becomes tedious and practically inapplicable [2]. Given the limits of manual indexing, the time and performance requirements, some documentary functions such as manual indexing must be automated.

IV. Automatic indexing

With the advent of computers, researchers have realized that they can use automatic techniques and software methods to index a collection of documents in order to facilitate Information searching and obtain precise results with a reduced time and resources. Several factors have encouraged *computer scientists*, library and information science researchers to find new automatic methods who are trying to enrich or replace manual indexing. The automation of indexing has helped to overcome the limits and inadequacies of intellectual indexing approaches such as cost and subjectivity. Unlike human indexing, automatic indexing uses a free vocabulary formed by extracting key terms (a single word or a group of words) characterizing documents.

Many statistical and / or linguistic indexing methods have been proposed to automatically extract the representative terms of a document:

1. Statistical indexing methods

Statistical methods of automatic indexing are based on purely mathematical and statistical calculations in order to define the weight of a word, according to different criteria such as:

- **Word frequency:** the weight of a word is calculated according to its number of occurrences (how many times a word appears in a document). the most frequent words in the document are the most significant and will serve as a descriptor. We can eliminate unimportant words (i.e. stop words, grammatical words). Stop words are basically a set of commonly used words in any language. Determinants, pronouns, prepositions, conjunctions, grammatical adverbs are stop words. We should rather focus on the important words (i.e. content words, open-class words and lexical words) those that have meaning. Nouns, adjectives, verbs and adverbs are content words.
- **Word density:** the density of a word is calculated according to the ratio between its occurrence in the document and the size of this document.
- **Word position in document:** the word position in document can have an influence on its weighting. For example, the position of the word in the title is more advantageous than at the end of the document.
- **Word writing style:** give the advantage to words in capital letters and in bold in the weighting.
- Etc.

In information retrieval, there are a multitude of similarity measures in the literature. The best-known are TF-IDF (Term Frequency - Inverse Document Frequency) [3], Dice similarity [4], Jaccard similarity [5], character n-gram similarity [6], Hidden Markov models [7], levenshtein distance [8] and Jaro-Winkler measure [9].

2. Linguistic methods

Linguistic methods of automatic indexing are a subdomain of Natural Language Processing (NLP). NLP is a multidisciplinary field that combines linguistics, computer science, information engineering, and artificial intelligence. These methods use different levels of analysis:

(i) The morphological analysis is made up of three steps:
   1. Segmenting (*Tokenization)* the text into sentences. A sentence is a character string located between a capital letter and a strong punctuation mark: full stop (period or full point), question mark and exclamation mark.

The full stop as a sentence separator can present ambiguities. It can be an abbreviation marker or titles prefixing the name of a person (e.g. Mr, Mrs, Mrs, Dr, etc.), part of an acronym (e.g. I.S.K.O.), etc.

2. *Segmenting sentences into words. A word is a single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence.* The separators are spaces, numbers, and weak punctuation marks (usually: comma, semicolon, colon, parentheses, ellipsis is also called a *suspension point*, dash, brackets and quotation marks).
3. Lexical analysis is composed on lexical and inflectional morphological analysis:

a. Lexical morphological analysis consists in studying the form of words which can be simple, complex (compounds), variable (nouns, verbs, determiners, pronouns and qualifying adjectives) or invariable (adverbs, prepositions and coordination conjunctions).
b. inflectional morphological analysis consists in studying the variation of lexical units as a function of grammatical factors. it represents the relationship between the different parts of a sentence and can concern a verb (conjugation) or a nominal group which depends on its grammatical category, its genre and its number.

(ii) The syntactic analysis (or parsing) allows to highlight the syntactic structure of a sentences by explaining the dependency relations between words. The purpose of this phase is to represent the structure of sentences using syntax trees. Syntactic analysis identifies syntactic groups such as noun phrases, verb phrases, prepositional phrase etc. These phrase groups are the basis of several indexing approaches [10, 11, 12].

3. Semantic indexing

The problem of indexing documents by words, or groups of words, is not using semantic relationships between descriptors such as synonymy, homonymy, polysemy relationships, etc. With the emergence of terminological resources such as ontologies [13], semantics has become a major challenge to consider. Semantic indexing [14, 15] uses the concepts and their relationships to represent documents and queries.

4. Social indexing:

Social indexing is a Web 2.0 technology, *also known as* social tagging, collaborative tagging, collaborative indexing *social* classification and Folksonomy. It involves a community of users freely creating and managing personalized tags (is a keyword or term) assigned to a web resource for the purposes of collaborative categorization and classification. "User are also actively involved in content creation, feedback and enrichment" [16]. Social indexing allows shared content collaborative enrichment web and creating new communities. There are several research studies on social indexing such as recommendation in social networks [17, 18, 19], improving information retrieval [20], information monitoring [21], etc.

V. Indexing methods for Big Data and Open Data

The rise of Big Data (or massive data, huge data) has followed the evolution of data storage and processing systems, notably with the advent of the *technology* of *cloud computing (*virtualization) and supercomputers. Big Data is also data but with a huge size**.** Big Data is a term used to describe a heterogeneous data sets that is huge in volume and yet growing exponentially with time. These *data* sets are *so* voluminous and complex that none of the traditional data management tools are able to store and manage them efficiently. Doug Laney [22] uses 3 properties or dimensions to define Big Data usually *called* the 3 Vs of Big Data (volume, variety and velocity). Volume refers to the *growing volume* of *data* generated through social media, websites, portals, online applications and *connected objects (smart objects).* Variety refers to the many types of data that are available which can be structured, semi-structured or unstructured, such as text, audio, pictures and video (i.e. multimedia documents). Multimedia documents require additional preprocessing and a classification of the incoming data into various categories. Velocity refers to the speed with which data are being generated, received, stored, processed, analyzed and exploited in real time. Big Data comes from various sources, such as published content on Internet, messages exchanged on social media, data transmitted by connected objects, climate data, demographic data, scientific and medical data, data from sensors, e-commerce transactions, company data, etc.

Open Data is an important source of data, it refers to digital data whose access, use, **re-use and Redistribution** (sharing) are public and free of rights (there should be no discrimination against persons or groups). They can be of public or private sector, produced and published by the government, a public service, a community or by a company. The exploitation of this data offers numerous opportunities and new perspectives to improve the performance of companies and to extend human knowledge in many fields.

The huge volume of data, the variety of structures and types of documents (text, image, sound and video) from heterogeneous sources are the biggest indexing problems. To overcome these problems, all indexed documents must be stored in the same format. The NoSQL (Not only Structured Query Language) databases [23, 24] are flexible and increasingly used with the rise of Big data to improve the performance of processing and analysis of distributed data. These data can have variable data structures different from those used by default in traditional *relational databases*. NoSQL databases do not use the *rows/columns/table* format. The most common types of *NoSQL databases* are *key-value, wide column*, document *and graph*:

- *Key-Value store:* A *key-value database, or key-value store*, is a *data storage* paradigm designed for *storing data in unique key-value pairs* where *each key is associated only with one value in a collection (*text, photo, video, object structure, …).
- *Wide-column store:* wide-column *databases* are designed for storing data as sections of columns where *each key is associated only with* a set of columns.

- Document store: Document databases use common notation formats like JavaScript Object Notation (JSON)or Extensible Markup Language (XML) to store documents. E*ach key is associated* a collection of key-value pairs stored in documents. This type of database is used to store structured and semi-structured documents.
- Graph Store: Graph database use graph theory to model data with nodes (entities or objects) and relationships (edges). Both nodes and relationships can have properties. This database type can store and analyze complex, dynamic and interconnected data. Many emerging problems such as social networks analysis, network routing, trend prediction, product recommendation, fraud detection can be represented using graph models and solved using graph algorithms [25].

VI. Conclusion

Indexing is a process used to extract descriptive elements from documents and users' requests. the aim of indexing is to improve searching for information by finding relevant documents in a collection of documents in a *reduced a search time.* Several studies have been developed to propose indexing approaches and methodologies ranging from manual and automatic methods to the emerging indexing methods for Big Data. This variety of methods must adapt and take advantage of *continuous technological evolution. In recent years, the emergence of Big Data, Open Data and* No SQL *databases have opened a new technological era and new research areas.* The purpose of these indexing methods is to allow the exploitation of huge digital data daily *produced* by humans and connected objects.

References:

[1] CHAUMIER Jacques, DEJEAN Martine, L'indexation documentaire, del'analyse conceptuelle à l'analyse morphosyntaxique, Documentaliste, vol.27, n°6, novembre-décembre 1990, pp.275-279.

[2] CLAVEL Geneviève, Walther Frédéric, WALTHER Joëlle, Indexation automatique de fonds bibliothéconomie, ARBIDO-R8, 1993, pp.14-19

[3] Salton G. et McGill M. J. (1983). Introduction to Modern Information Retrieval. McGraw Hill Book Co.,New York, 1983.

[4] Sneath, P. H. et Sokal, R. R. (1973). Numerical Taxonomy - The Principles and Practice of Numerical Classification. San Francisco: W. H. Freeman and Company.

[5] Grefenstette G. (1994). Exploration in Automatic Thesaurus Discovery, Londres, Kluwer Academic Publishers.

[6] Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal,27(3), 379–423.Continued in the following volume.

[7] Baum, Leonard E., and Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." *Annals of Mathematical Statistics* (1966): 1554–1563.

[8] Levenshtein, V. I. (1965), "Binary codes capable of correcting deletions, insertions, and reversals.", Doklady Akademii Nauk SSSR, 163 (4): 845–848

[9] [n+3] Jaro, Matthew A. (1989) "Advances in Record-linkage Methodology a Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, 89, pp. 414-420.

[10] Jean-Pierre Chevallet, Hatem Haddad (2001). Proposition d'un modèle relationnel d'indexation syntagmatique : mise en œuvre dans le système iota. *INFORSID 2001*, Genève-Martigny.

[11] Sahbi SIDHOM (2002). Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances. Thèse de doctorat de l'Université Claude Bernard - Lyon Ie 11/03/2002 (France).

[12] Zoulikha BELLIA HEDDADJI (2005). Modélisation et classification de textes. Application aux plaintes liées à des situations de pollution de l'air intérieur. Thèse de doctorat de l'Université Paris Descartes (France).

[13] Jonquet, Clement & Coulet, Adrien & Shah, Nigam & Musen, Mark. (2010). Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical. 21èmes Journées Francophones d'Ingénierie des Connaissances.

[14] Abdelkader Hamadi (2014). Utilisation du contexte pour l'indexation sémantique des images et vidéos. Intelligence artificielle [cs.AI]. Université de Grenoble, France.

[15] Ameni Yengui. Système de recherche d'information sémantique pour les bases de visioconférences médicales à travers les graphes conceptuels. Recherche d'information (2016) [cs.IR]. Faculté des sciences économiques et de gestion Sfax (Tunisie).

[16] Rückemann, C.P. (2012), « Integrated Information and Computing Systems for Natural, Spatial, and Social Sciences", Information Science Reference, 543 p, ISBN: 1466621915, 9781466621916

[17] Mohammed Ryadh Dahimene (2014). Filtrage et Recommandation sur les Réseaux Sociaux. Thèse de doctorat ÉCOLE DOCTORALE INFORMATIQUE, TÉLÉCOMMUNICATION ET ÉLECTRONIQUE- PARIS 8 Décembre 2014 (France).

[18] Mohamed Nader Jelassi, Sadok Benyahia, Mephu Nguifo Engelbert (2016). Étude du profil utilisateur pour la recommandation dans les folksonomies. IC2016 : Ingénierie des Connaissances, Jun 2016, Montpellier,France.

[19] Samia Beldjoudi, Hassina Seridi, Abdallah Benzine (2016). Améliorer la Recommandation de Ressourcesdans les Folksonomies par l'Utilisation de Linked Open Data. IC2016 : Ingénierie des Connaissances,Jun 2016, Montpellier, France.

[20] Ismail Badache (2016). Recherche d'information sociale : exploitation des signaux sociaux pour améliorer la recherche d'information. Université Paul Sabatier - Toulouse III, France.
[21] Pirolli, Fabrice. (2011). Pratiques d'indexation sociale et démarches de veille informationnelle. Études de communication. 53-66. 10.4000/edc.2615.
[22] Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6.
[23] Moniruzzaman, A B M & Hossain, Syed. (2013). NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. Int J Database Theor Appl. 6.
[24] Bathla, Gourav & Rani, Rinkle & Aggarwal, Himanshu. (2018). Comparative study of NoSQL databases for big data storage. International Journal of Engineering & Technology. 7. 10.14419/ijet.v7i2.6.10072.
[25] Skhiri, S., & Jouili, S. (2013). Large Graph Mining: Recent Developments, Challenges and Potential Solutions. InBusiness Intelligence (pp. 103-124).