From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art
Nabil Khemiri, Sahbi Sidhom

To cite this version:
Nabil Khemiri, Sahbi Sidhom. From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art. Multi-Conference OCTA’2019 on: Organization of Knowledge and Advanced Technologies, Université de Tunis; ISKO-Maghreb Chapter, Feb 2020, Tunis, Tunisia. pp.171-175. hal-02933709

HAL Id: hal-02933709
https://hal.science/hal-02933709
Submitted on 8 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art

Nabil Khemiri, Sahbi Sidhom

To cite this version:
Nabil Khemiri, Sahbi Sidhom. From manual indexing to automatic indexing in the era of Big Data and Open Data: a state of the art. Multi-Conference OCTA’2019 on: Organization of Knowledge and Advanced Technologies, Université de Tunis; ISKO-Maghreb Chapter, Feb 2020, Tunis, Tunisia. pp.171-175. hal-02933709

HAL Id: hal-02933709
https://hal.archives-ouvertes.fr/hal-02933709
Submitted on 8 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain
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.

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 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.

![Information Search System](image)

**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 a 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. Each 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 databases 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 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:


