Revista Española de Documentación Científica
45(4), octubre-diciembre 2022, e338
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2022.4.1917

ESTUDIOS / *RESEARCH STUDIES*

Automatic indexing of scientific articles on Library and Information Science with SISA, KEA and MAUI

Isidoro Gil-Leiva*, Pedro Díaz Ortuño*, Renato Fernandes Corrêa**

* Universidad de Murcia
Correo-e: isgil@um.es | ORCID iD: https://orcid.org/0000-0002-7175-3099
Correo-e: diazor@um.es | ORCID iD: https://orcid.org/0000-0002-2975-766X
** Universidade Federal de Pernambuco
Correo-e: renato.correa@ufpe.br | ORCID iD: https://orcid.org/0000-0002-9880-8678

Recibido: 02-09-21; 2ª versión: 26-10-21; Aceptado: 02-11-21; Publicado: 18-10-2022

Cómo citar este artículo/Citation: Gil-Leiva, I.; Díaz Ortuño, P.; Fernandes Corrêa, R. (2022). Automatic indexing of scientific articles on Library and Information Science with SISA, KEA and MAUI. *Revista Española de Documentación Científica*, 45 (4), e338. https://doi.org/10.3989/redc.2022.4.1917

Abstract: This article evaluates the SISA (Automatic Indexing System), KEA (Keyphrase Extraction Algorithm) and MAUI (Multi-Purpose Automatic Topic Indexing) automatic indexing systems to find out how they perform in relation to human indexing. SISA's algorithm is based on rules about the position of terms in the different structural components of the document, while the algorithms for KEA and MAUI are based on machine learning and the statistical features of terms. For evaluation purposes, a document collection of 230 scientific articles from the *Revista Española de Documentación Científica* published by the Consejo Superior de Investigaciones Científicas (CSIC) was used, of which 30 were used for training tasks and were not part of the evaluation test set. The articles were written in Spanish and indexed by human indexers using a controlled vocabulary in the InDICES database, also belonging to the CSIC. The human indexing of these documents constitutes the baseline or golden indexing, against which to evaluate the output of the automatic indexing systems by comparing terms sets using the evaluation metrics of precision, recall, F-measure and consistency. The results show that the SISA system performs best, followed by KEA and MAUI.

Keywords: automatic indexing; automatic indexing systems; SISA; KEA; MAUI; indexing assessment.

Indización automática de artículos científicos sobre Biblioteconomía y Documentación con SISA, KEA y MAUI

Resumen: Este artículo evalúa los sistemas de indización automática SISA (Automatic Indexing System), KEA (Keyphrase Extraction Algorithm) y MAUI (Multi-Purpose Automatic Topic Indexing) para averiguar cómo funcionan en relación con la indización realzada por especialistas. El algoritmo de SISA se basa en reglas sobre la posición de los términos en los diferentes componentes estructurales del documento, mientras que los algoritmos de KEA y MAUI se basan en el aprendizaje automático y las frecuencia estadística de los términos. Para la evaluación se utilizó una colección documental de 230 artículos científicos de la *Revista Española de Documentación Científica,* publicada por el Consejo Superior de Investigaciones Científicas (CSIC), de los cuales 30 se utilizaron para tareas formativas y no formaban parte del conjunto de pruebas de evaluación. Los artículos fueron escritos en español e indizados por indizadores humanos utilizando un vocabulario controlado en la base de datos InDICES, también perteneciente al CSIC. La indización humana de estos documentos constituye la referencia contra la cual se evalúa el resultado de los sistemas de indización automáticos, comparando conjuntos de términos usando métricas de evaluación de precisión, recuperación, medida F y consistencia. Los resultados muestran que el sistema SISA funciona mejor, seguido de KEA y MAUI.

Palabras clave: indización automática; sistemas de indización automática; SISA; KEA; MAUI; evaluación de indización.

1. INTRODUCTION

Document production has grown exponentially since the 1950s. Not only has publication increased significantly but increasing numbers of documents are processed and disseminated, giving rise to a need for more efficient and faster information processing systems. For instance, in large bibliographic databases such as Scopus, some three million documents are incorporated each year. Furthermore, libraries are integrating large quantities of electronic books, papers, theses and dissertations, which they are unable to process adequately in order to make them accessible through a catalogue or institutional repository. In the documentary management of the aforementioned information systems, the indexing of content to facilitate access plays a fundamental role.

ISO standard 5963-1985 defines indexing as "The act of describing or identifying a document in terms of its subject content". To this it may be added that, on occasion, concepts are normalized and controlled by controlled vocabulary, as otherwise it would be natural language indexing, and likewise that indexing is carried out – be it consciously or unconsciously – according to the users' information needs in order to convert these (in natural or controlled language) into a search query. Hence, indexing constitutes an essential process for storing documents and may also be so for retrieving information if the result of the indexing (keywords, descriptors, subjects indexing) is used later for retrieval. Indexing is therefore the cornerstone of document management systems as it is essential to represent the contents of documents and to facilitate their subsequent retrieval.

Programs to increase workflow performance were first created in the late 1950s and the terminology used in the literature to refer to the process of making indexing automatic is varied, although the most used term is "automatic indexing". The definition of automatic indexing can derive from three perspectives: a) computer aided indexing during storage; b) semi-automatic indexing; and c) automatic indexing (Gil-Leiva, 2008: 320). From the 1970s to the present, several automatic indexing programs have been developed. Without intending to be exhaustive, some of them are:

- MAI (Klingbiel, 1973; Silvester et al., 1994). Machine-aided indexing (MAI) was developed at the US Defense Documentation Center.
- The SMART project by Salton (1989, 1991). One of the first to incorporate advances produced by the automatic processing of natural language, it brings in tools to extract word roots, thesaurus, morphological or syntactic analyzers.
- CLARIT (Evans, 1990; Evans et al., 1991a; Evans et al., 1991b). CLARIT (Computational-Linguistic Approaches to Indexing and Retrieval of Text) used a lexicon for general English which consists of approximately 100,000 root forms and hyphenated phrases, tagged for syntactic category and irregular morphological variation; a morphological analyzer; a lexical disambiguator; a noun phrase grammar; and various indexing algorithms, such as the ranking indexing terms.
- SIMPR (Karetnyk, Karlsson & Smart, 1991). In the SIMPR (Structured Information Management: Processing and Retrieval) project, output from a morphological analyzer was used to provide the input to indexing software, from which some indexing terms are finally obtained, and are validated manually.
- SAPHIRE (Hersh & Greenes, 1990; Hersh et al., 1991). SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment) is a project in the field of biomedicine and the Medline database.
- Indexing Initiative/Text Indexer-MTI (Humphrey & Miller, 1987; Humphrey, 1999; Humphrey et al., 2006; Aronson et al., 2000; Mork et al., 2017). Projects whose mission among library teams is to explore indexing methodologies to ensure the quality and currency of the document collections of the US National Library of Medicine (NLM). The NLM Medical Text Indexer (MTI) is the core product of this project and has been providing automated indexing recommendations since 2002 to the present day.
- CAIT, Luxid, AgNIC (Irving, 1997; Salisbury & Smith, 2014). At the United States National Library of Agriculture, various projects have been implemented, including CAIT (Computer-Assisted Indexing Tutor), a program which sought to enhance the quality of indexing and the training of new indexers (Irving, 1997). This was followed by the acquisition of the Luxid indexing software from the TEMIS company in 2011 to considerably increase the production capacity of the six indexers from 75,000 articles a year to about 300,000. In addition, more recently there was AgNIC (Agriculture Network Information Collaborative), which bases automatic indexing on the Thesaurus, as Luxid does.
- CISMeF (Chebil et al., 2012). The *Catalogue et Index des Sites Médicaux de langue Française*, a system implemented for the automatic indexing of medical information resources.
- Annif (Suominen, 2019). Developed by the National Library of Finland, Annif is an open source tool and microservice for automated subject indexing based on a combination of existing nat-

ural language processing and machine learning tools, combining multiple approaches and existing open source algorithms.

The scientific literature on automatic indexing is extensive and mainly devoted to describing and evaluating prototypes. There is a long tradition of evaluation and competition of systems and prototypes, starting in the 1960s with the Cranfield experiments. Subsequently, from 1992 onwards, the Text Retrieval Evaluation Conference (TREC) initiative began to gain momentum, becoming the largest annual competition centring on information retrieval. The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as the Cross-Language Evaluation Forum) also emerged from this event and also promotes innovation and the development of information access systems in Europe on an annual basis. A third important evaluation-competition initiative is SemEval (Semantic Evaluation), which has focused on the meaning of language since 1998. Each of these three initiatives has featured spaces dedicated to indexing and automatic indexing. In a way, the work presented here is imbued with the spirit of the above initiatives in the sense of an evaluation-competition based on a test set, golden indexing and metrics.

Furthermore, since the early 2000s, scientific publications have been impregnated with the Semantic Web (which uses XML [eXtensible Markup Language], RDF [Resource Description Framework] and OWL [Ontology Web Language]), with the addition in 2012 of the Journal Article Tag Suite (JATS) standard for describing the formal, textual and graphic content of scientific articles. According to the 2020 Scholastica report, 35% of scientific journal publishers are already using the XML or XML JATS format. The *Revista Española de Documentación Científica* itself in 2013 extended its publication format from PDF to HTML and XML. The Redalyc platform has been using XML JATS extensively since 2016. Thus, this combination of Open Science and Semantic Web has given rise to the so-called semantic publishing that facilitates dissemination, exchange, reuse of information, retrieval and automatic information processing.

The work presented here uses the web version of SISA, which represents a significant advance over the previous version in Windows. In some way, SISA imitates the work of a human indexer during the reading phase of the documents. Besides, this approach, based on the position occupied by terms in the documents and aligned with the increasing number of scientific articles published using XML JATS, makes SISA a different system to other automatic indexing prototypes. KEA and MAUI, for example, use a machine learning model that starts from a training set with documents and their indexing terms to then predict the indexing terms for new documents using a Naïve Bayes classifier and Decision Tree classifier respectively. The machine learning model requires input consisting of statistical features of terms in a document, and its output comprises weighting of terms. A trained indexing model performs the term-weighting process, based on calculating characteristics for each term, such as occurrence, position, size, the probability of being a keyword and semantic relations. KEA uses the two first feature types, while MAUI can use all of them.

Therefore, the main objective of this article is to find out how the SISA algorithm (based on the position that the terms occupy in the different parts of the document) performs compared to the KEA and MAUI algorithms, which are machine-learning based. This work may also enable further improvement of SISA through the feedback obtained from studying and analysing the human indexing of the InDICES database, which for the purposes of this study has been taken as a benchmark, as well as from analysing the outputs of KEA and MAUI. Thus, in addition to highlighting the importance of the methodology on which SISA is based compared to other indexing systems, this research seeks to answer specific questions such as: what is the average time taken to index a scientific article in SISA, KEA and MAUI compared to a human indexer?; what is the average number of terms assigned by these systems compared to human indexing?; and what maximum and average F-measure and consistency indices can these automatic indexing systems achieve compared to human indexing?

2. MATERIALS AND METHOD

Our study consists of six main parts: a) selection of the prototypes to carry out the comparison-evaluation with SISA; b) collection of the sources and data to carry out the evaluation (document collection, controlled vocabulary and human-assigned golden indexing); c) selection of the evaluation metrics; d) configuration and indexing of the document collection using the three prototypes; e) application of the evaluation metrics to obtain the precision, recall, F-measure and consistency indices; f) analysis and discussion of the results. We provide more details of these below.

2.1 Automatic indexing systems

SISA, KEA and MAUI are three automatic information processing tools specifically for automatic document indexing. The reasons that led to KEA and MAUI being selected include the following: all three have their origins in academic settings; they are free, open-source software tools; they all have in common the option of performing automatic indexing by assignment, which means terms being assigned from a given controlled vocabulary; and, finally, KEA and MAUI use an automatic indexing model based on a machine learning algorithm and statistical features of terms.

As has been pointed out above, the evaluation-comparison of these three indexing tools will make it possible to examine in depth two different approaches to the automation of indexing, and therefore to find out which of them is closer to human indexing. One approach is based on the position that terms occupy in the different parts of documents (keeping in mind here that the ISO standard for document indexing itself indicates to which parts of documents indexers should pay attention when extracting and assigning terms) while the other is a machine learning approach (based on a model which has been trained through a small collection of documents and the golden indexing of those documents to predict the indexing terms for a new set of documents). The three automatic indexing systems are presented below.

SISA

The conceptual development of SISA began in the mid-1990s. SISA is an automatic indexing system developed in JAVA to extract information from documents. It performs automatic indexing of scientific articles, legislation (laws, decrees) and judicial sentences, although to date the document typology it works with most is scientific articles. It processes documents in TXT, HTML or XML formats. It also uses a controlled vocabulary in TXT or SKOS format.

SISA is available on the Internet to users through a password (http://fcd1.inf.um.es:8080/portal/). SISA processes documents written in Spanish, Portuguese and English. It processes texts with tags that indicate each of the components of the documents depending on document typology: for example, for articles it processes title, abstract, keywords, headings, first paragraph, table title, chart title, conclusions and references. Currently, it directly processes XML articles from the *Revista Española de Documentación Científica* and HTML articles derived from the JATS standard of the Redalyc platform. If the documents do not contain the necessary marks, the software has a wizard to assist with labelling.

These baseline fundamentals of SISA appear to align with the current development of the JATS standard for scientific articles. Marks and a set of rules based on heuristic (positioning – titles, abstracts, author keywords, headings, first paragraphs of headings, table titles, graph titles, conclusions and references) and statistical methods (frequency, TF-IDF) are the hallmark of this software.

Ever since its first version in 2002, SISA has offered the option of processing using automatic indexing or semi-automatic indexing. With semi-automatic indexing, users can edit the proposed indexing for each document through a wizard that shows each indexing term or phrase in its context. The successive tasks for automatic indexing of a document are the following: labelling the documents or uploading pre-marked documents; processing (applying stemming, calculating the TF-IDF and IDF and recording the place where terms and phrases appear); and then indexing the documents, taking into account the activated rules. It also includes an evaluation module that measures recall, precision and F-measure in information retrieval (Gil Leiva, 2008, 2017a).

SISA has been used in several experiments. Lima and Boccato (2009) used SISA to evaluate the performance of descriptors in manual, automatic and semi-automatic indexing processes. In the study by Souza-Rocha and Gil-Leiva (2016), a comparison was performed of the indexing of the same test set by SISA and by the PyPLN platform developed at the Getulio Vargas Foundation’s School of Applied Mathematics in Rio de Janeiro, Brazil. Available functions of the platform include part-of-speech tagging, word and sentence level statistics, n-grams extraction and word matching. Gil-Leiva (2017a) investigated the capabilities of SISA by comparing its indexing of a test set of scientific articles on Agriculture with their indexing by the Agricola, WOS and SCOPUS databases. Later work by Gil-Leiva (2017b) aimed to determine which rules (rules of position of the terms in the document or TF-IDF rules) provide the best indexing terms, using SISA to obtain the automatic indexing of 200 scientific articles on fruit growing written in Portuguese. In addition, more recently the indexing of the desktop version of SISA has been compared with MAUI (Silva & Correa, 2020; Silva et al., 2020).

The work presented here uses the web version of SISA, which represents a significant advance over the Windows version in terms of processes, functionalities, document input formats and the incorporation of statistical rules, but mainly due to the greater number of heuristic rules based on the position of the terms in the documents. At the current time, SISA can process tagged PDF, TXT and

XML documents published by the Revista Española de Documentación Científica and XML JATS articles produced and published by the editors of Redalyc.

KEA

The Keyphrase Extraction Algorithm (KEA) is a project developed by the "Digital Library" and "Machine Learning" research groups at the University of Waikato. New Zealand Digital Library Project members have developed a range of practical software packages in the course of their research. The home page is available at http://community.nzdl.org/kea/.

KEA is an algorithm for automatically extracting keyphrases from text documents, assuming that keyphrases provide semantic metadata that summarize and characterize documents. Implemented in Java, it is platform-independent and an open-source software distributed under the GNU General Public License. The system is simple, robust and publicly available. It can be used either for free indexing or for indexing with a controlled vocabulary (Thesaurus, Subject Headings). The KEA website provides examples of its use in different domains with a thesaurus: the AGROVOC thesaurus, Medical Subject Headings (MESH) thesaurus and High Energy Physics thesaurus (HEP).

KEA includes a machine learning component. In experiments using KEA, it is utilized to split the document collection into a training and a test set. The training set is applied to train a machine learning model to classify candidate keyphrases. The test set makes it possible to evaluate the effectiveness of KEA in terms of how many author-assigned keyphrases are correctly identified.

KEA performs the following processing steps: it identifies candidate keyphrases using lexical methods; it calculates the feature values for each candidate; and it uses a machine learning algorithm to predict which candidates are good keyphrases. The machine learning scheme first builds a prediction model using training documents with known keyphrases, and then it uses the model to find keyphrases in new documents (Witten et al., 1999). The machine learning model is constructed automatically from these labelled training examples using the WEKA machine learning workbench. KEA (Frank et al., 1999) uses the Naïve Bayes classifier, which implicitly assumes that the features are independent of each other.

Medelyan (2005) extended KEA into a new version called KEA++. The indexing process is extended through analyzing semantic information about a document's terms, i.e. the relationship between the terms and their relationship to other terms in the thesaurus. It combines two approaches into a single process whereby terms and phrases are extracted from documents (keyphrase extraction, phrases analyzed to select the most representative ones) but need to be part of a controlled vocabulary (assignment of keyphrases, documents are classified into a pre-defined number of categories corresponding to descriptors).

As KEA++ is a supervised learning approach, it involves two phases: training and testing. It builds a learned model by applying a Naive Bayes algorithm using training data labelled with thesaurus terms. The extraction phase uses the learned model to identify, from a thesaurus, the most significant keyphrases based on certain properties (features) and assign them to the test documents. After computing the feature values for the training set, the model built is used to extract keyphrases from new documents. Each candidate keyphrase is marked as a positive or negative example, depending on whether users have assigned it as a keyphrase or tag to the corresponding document (Medelyan, 2009).

KEA has been used in a number of different experiments. El-Haj et al. (2013) experimented with KEA in a different domain to the examples available on the project website, when they used KEA to examine the quality of the automated indexing process based on a controlled vocabulary called the Humanities and Social Sciences Electronic Thesaurus (HASSET). Khan et al. (2011) and Irfan et al. (2014) set out to improve the functioning of KEA++ based on more efficient exploitation of the existing hierarchical relationships in the domain vocabularies used by the system. Wang et al., (2015) designed DIKEA with the idea of improving the performance of KEA++ in two ways: by extracting keywords from documents and by not depending on a specific vocabulary. Other experiments in which KEA has been used include those by Duwairi and Hedaya (2016) for documents with Arabic news, those by Akhtar et al. (2017) to create a hierarchy of keywords extracted from documents and the research by Gopan et al. (2020), who compared the keyword extraction algorithms of KEA, TextRank and PositionRank.

MAUI

Olena Medelyan developed the MAUI (Multi-purpose Automatic Topic Indexing) as part of her doctoral project, under the supervision of Ian H. Witten and Eibe Frank, at the Department of Computer Science at the University of Waikato, New Zealand, in 2009.

To represent the content of the documents, the software uses statistical and linguistic methods for extracting and weighting terms. The software extracts terms, or generates candidates, by identifying n-grams not containing punctuation marks and not beginning or ending with stopwords, after normalization/conflation by stemming. A trained supervised indexing model performs the term-weighting process, based on the calculation of characteristics for each term, such as occurrence, position, size, the probability of being a keyword and semantic relations.

The software allows users to perform the following tasks (Medelyan, 2009): assigning terms with a controlled vocabulary or thesaurus; subject indexing; topic indexing with Wikipedia terms; keyword extraction; terminology extraction; automatic markup, terminology extraction and semi-automatic topic indexing; and keyword extraction from the text using a controlled vocabulary as a source of terms.

MAUI uses a machine learning algorithm to generate a model for selecting index terms based on the intellectual indexing of a set of documents. For this reason, it requires input consisting of a training set, made up of documents and respective terms for human indexing.

To train the indexing model, the data entry requirements for MAUI are the following:

- Thesaurus in SKOS format: a text file containing the authorized terms and their non-preferred terms, as well as the semantic relations between authorized terms;
- Stemmer of words for the language used in the text of the documents: a software component that reduces the words to an approximation of their stem;
- Pre-established list of stopwords in the document text language: a text file containing words without thematic meaning for elimination, such as connectors and articles;
- A training set: a set of documents and respective indexing terms for training a machine learning model;
- All input files, including the texts for indexing, must be in text format with UTF-8 without BOM charset.

After the model has been trained, the data entry requirements for automatic indexing of a text document using MAUI processing are: the full text of the document to be indexed; the trained model; a list of stopwords; a stemmer; and a thesaurus or controlled vocabulary.

According to Medelyan (2009), MAUI performs indexing through the following processing steps:

1. Generation of candidate topics - extraction of candidate terms for indexing;
2. Calculation of characteristics - calculation of characteristics for candidate terms;
3. Construction of the indexing model - training of an indexing model considering the terms assigned by the indexers to each document in the training set;
4. Application of the learned model to select topics for other documents - application of the trained indexing model to propose indexing terms to other documents.

The quality of the training set provided for the machine learning algorithm is the key to better quality in automatic indexing, whether the indexing process involves automatically extracting keywords or assigning controlled vocabulary descriptors.

In addition to the experiments by Medelyan (2009), several other studies have applied MAUI. For instance, MAUI was tested to extract keyphrases from scientific texts in English at the SEMEVAL 2010 conference (Kim et al., 2013), while Mynarz and Škuta (2010) used MAUI together with other applications to implement an automatic indexing system for Czech grey literature and Sinkkilä, Suominen and Hyvönen (2011) tested MAUI and three different stemming and derivation tools on Finnish texts to see which obtained the best indexing terms. The report by Shams and Mercer (2012a) looked at the indexing performance of MAUI when paired with a text extraction procedure called text denoising, and MAUI was also applied to extract keyphrases from scientific texts written in Spanish (Aquino & Lanzarini, 2015). Silva, Correa and Gil-Leiva (2020) implemented MAUI to assign thesaurus terms to scientific texts written in Portuguese, in addition to comparing MAUI indexing with indexing from a desktop version of SISA. In the research by Gopan et al. (2020), MAUI's keyword extraction algorithm was compared with KEA, TextRank and PositionRank.

2.2 Document collection

A document collection was created, made up of 230 scientific articles published in the *Revista Española de DocumentaciónCientífica* of the Consejo Superior de Investigaciones Científicas (CSIC– the Spanish National Research Council). This long-running journal began to publish its articles in PDF, HTML and XML formats in 2013, using an XML format with DTD NLM evolved in the current JATS standard. The criteria for selecting the 230 documents that made up the document collection were simple: articles published in the "Studies" section from 2013 to 2020 that were written in Spanish. To complete the collection, the "Note and experiences" section, which publishes scientific articles of similar relevance and formal structure, was also included.

The training set for KEA and MAUI software comprised 30 randomly selected articles, since SISA does not require an initial machine learning phase for the system. The remaining 200 articles constituted the test set that was applied to evaluate the automatic indexing performed by SISA, KEA and MAUI and later, to carry out the comparative studies. SISA used the 200 articles in XML format while KEA and MAUI worked with the extracted text in TXT format.

2.3 Controlled vocabulary

A controlled vocabulary composed of 10,981 terms was used, of which 8,214 were preferred terms and 2,767 non-preferred terms. The controlled vocabulary was provided by those responsible for the InDICEs database of the Consejo Superior de Investigaciones Científicas. In this database, the controlled vocabulary is utilized for human indexing of scientific journals published in Spain, as well as conference proceedings on Library and Information Science. Therefore, the 230 articles that make up the test collection were indexed using this controlled vocabulary from 2013 to 2020.

2.4 Gold indexing

As golden indexing, we used the indexing assigned by the CSIC's indexers to each of the 200 documents that made up the test set. For this purpose, a query was made in the InDICEs database to retrieve all the documents published in the *Revista Española de Documentación Científica* from 2013 to 2020. Subsequently, the indexing assigned to each of the 200 documents was compiled and was used for the evaluation processes. It was also ascertained whether all the indexing terms assigned to the 200 documents in the test set were present in the controlled vocabulary. Various terms were identified which had been used as indexing terms but did not appear in the controlled vocabulary, and we proceeded to include them.

2.5 Evaluation metrics

The metrics used to compare indexing were of two types: on the one hand, precision, recall and F-measure, and on the other hand, the indexing consistency measure.

Precision, recall and F-measure are used extensively in the field of information retrieval (Gupta et al.,2015; Gil-Leiva, 2017a; Al-Zoghby, 2018; Seiler, Hübner & Paech, 2019; and Lin et al., 2020, to cite some recent examples). These metrics have been adapted for other purposes such as evaluating information extraction, summarizing and comparing automatic indexing with golden indexing.

In order to evaluate the systems' automatic indexing output, the output keywords (the generated keywords) were compared with the human-assigned keywords for each document. We considered an output keyword to be "relevant" if it was an exact match with a human-assigned keyword. Precision, recall and F1 scores were calculated at document level, and then aggregated over the document collection.

Examples of use of these metrics to compare indexing pairs include their use by Krapivin et al. (2008), Shams and Mercer (2012b) and El-Haj et al. (2013) to compare KEA results; by Bandim and Correa (2019) and Silva, Correa and Gil-Leiva (2020) to compare SISA; by Névéol et al. (2005) and Chebil et al. (2012) to compare CISMEF automatic indexing; and by Rae et al. (2021) to evaluate the Medical Text Indexer (MTI) system.

The indexing consistency measures applied are those proposed by Hooper (1965) and Rolling (1981), specifically designed to compare indexing. Experiments using Hooper's measure include those by Mork, et al. (2017) in their evaluation of Medical Text Indexer system and by Silva, et al. (2020) to evaluate SISA; however, Sinkkilä et al. (2011), for example, used Rolling's measure to evaluate MAUI.

In our experiment, we applied Hooper's measures to compare the indexing performed automatically by SISA, KEA and MAUI with human indexing (golden indexing).

2.6 Experiments

To apply KEA and MAUI, it was first necessary to convert the XML and PDF documents into text files (.txt files), as well as extract the descriptors of the human indexing and fill the key files (.key files).

Having compiled the golden indexing for the test set and following the training phase required by KEA and MAUI using 30 documents, the automatic indexing of the test set of 200 articles in XML and TXT format by the three tools was begun. The use of the same test set and controlled vocabulary for the automatic indexing of articles allows comparisons to be made between the automatic indexing of the three software programs.

We set a threshold of 10 as the maximum number of keywords that KEA and MAUI could generate. The threshold value was chosen to reflect the number of keywords that we estimated it would be reasonable for KEA and MAUI to extract from the documents in the test set, given the number of keywords assigned by the golden indexing.

SISA System configuration

There were nine rules used in this experiment. Seven were position rules, that is, the position of a term in the text is considered. As they are scientific articles, the following parts were used: title, abstract, keywords, section headings, first paragraph of each section, the other paragraphs of that section, table titles, graph titles, conclusions and references. The position rules utilize a single part of the text (for example, Rule 1 proposes terms present in the title and Rule 2 proposes terms present in the author's keywords) or are activated when a term appears in several parts at the same time (for example, Rule 3 proposes the term if it appears in all these parts: abstract, section headings, first paragraph, other paragraphs, conclusions and references). In addition, the experiment used two statistical rules: one on the TF-IDF (for instance, Rule 8 proposes terms with a TF-IDF of at least 0.018) and the other on the total frequency of occurrence of a term in the document (for instance, Rule 9 proposes terms with a minimum frequency of 40 occurrences in the text).

All the terms (or their synonyms) selected by the rules must be present both in the document and in the controlled vocabulary. Several rules can propose the same term. At present, in the set of terms proposed to index a document, there is no weighting of the primary descriptor or secondary descriptor type to try to give more or less importance to one descriptor over another.

SISA does not currently allow users to set or limit the number of indexing terms for each document.

KEA System configuration

The source code of KEA version 5.0 (KEA++) can be downloaded from https://code.google.com/archive/p/kea-algorithm/ and https://github.com/EUMSSI/KEA. For the experiments, KEA was downloaded from the latter URL for importation in Eclipse IDE.

The following KEA input parameters were specified via changes in source code: training set directory, test set directory, vocabulary controlled in SKOS format, Spanish language selection, selection of the Spanish Snowball Stemmer for stemming words of the Spanish language, selection of stopwords list for Spanish, UTF-8 encoding for files, minimum frequency of occurrence for candidate terms equals 1 (training) and proposed indexing terms for each document set to 10.

KEA was applied to the corpora in stemming mode using Spanish Snowball Stemmer.

MAUI System configuration

The source code of MAUI used in the experiments was downloaded from https://github.com/zelandiya/maui-standalone.

The following MAUI configuration was set via changes in source code: selection of the stemmer for the Spanish language provided by Apache Lucene; selection of stopwords list for Spanish coded in the MAUI software; selection of the types of characteristics for the candidate terms based on frequency, position and size.

The following parameter values were provided for MAUI in the training phase: training set directory, vocabulary controlled in SKOS format, Spanish language selection, UTF-8 encoding for files and minimum frequency of occurrence for candidate terms equals 2.

The following parameter values were provided for MAUI in the test phase: the indexing model, test set directory, vocabulary controlled in SKOS format, Spanish language selection, UTF-8 encoding for files, the number of terms equals 10 and probability threshold equals 0.05.

3. RESULTS AND DISCUSSION

3.1 Processing time

Garcia Gutierrez (1984: 115) cited an experiment carried out in the early 1970s to learn about the reality of indexing in Great Britain, in which about twelve minutes were spent to obtain eleven to twenty keywords; regarding the findings of Garcia Gutierrez, the present study considers that the time taken to convert keywords in natural language into descriptors of a thesaurus or controlled vocabulary should be added. However, Farrow (1994: 158), citing Cleverdon (1962), indicated that the optimum time for indexing technical reports could be four minutes plus 60%, depending on the working conditions. On the other hand, Amat (1989: 176) stated that an average time of twenty minutes would be required to obtain about ten terms. And finally, in the AusLit repository they recommend using the following benchmarks: novel/drama, 30 minutes; biography/autobiography, 30 minutes; short story, 10-20 minutes; verse, 5 minutes; and critical article, 20-30 minutes. We considered that the indexing of a scientific article could take between ten and fifteen minutes, to try to establish some correlation between human and automatic indexing (Table I).

Table I: Estimated processing times for items.

	one article	**200 articles**
Manual indexing	10-15 minutes	33-50 horas
Automatic indexing by SISA	8.1 seconds	1620 seconds (27 minutes)
Automatic indexing by KEA	0.45 seconds	90 seconds
Automatic indexing by MAUI	0.1 second	10 seconds

Among the automatic indexing systems, SISA takes by far the longest time to process documents. However, for all the systems, processing times are much shorter than the processing time required for human indexing.

3.2 Mean number of indexing terms assigned

Graph 1 shows the mean number of indexing terms assigned in human indexing (golden indexing). In addition, for each automatic indexing system, it shows the average number of indexing terms assigned as well as the average number of terms in common with the terms assigned during golden indexing. Here we can see that KEA and MAUI have similar average values. SISA has an average number of assigned terms per document that is closer to the average number of golden indexing terms and an average number of terms shared with golden indexing per document almost one unit higher than KEA and MAUI.

3.3 Precision, recall and F-measure values

The classic formulas used to calculate the measurements in the experiment are shown in Table II, as well as the results obtained for article number 4 by the three automatic indexing systems.

In Table II, the D column value is the identifier of a document in the test set; the C column value is the number of correct terms assigned by each system in comparison to golden indexing for a document; the A column value is the total number of terms assigned by each system for a document; the GI column value is the number of golden indexing terms for a given document; the P column value is the precision; the R column value is the recall; and the F1 column value is the F-measure obtained by each system for a document.

Graph 1: Assigned terms

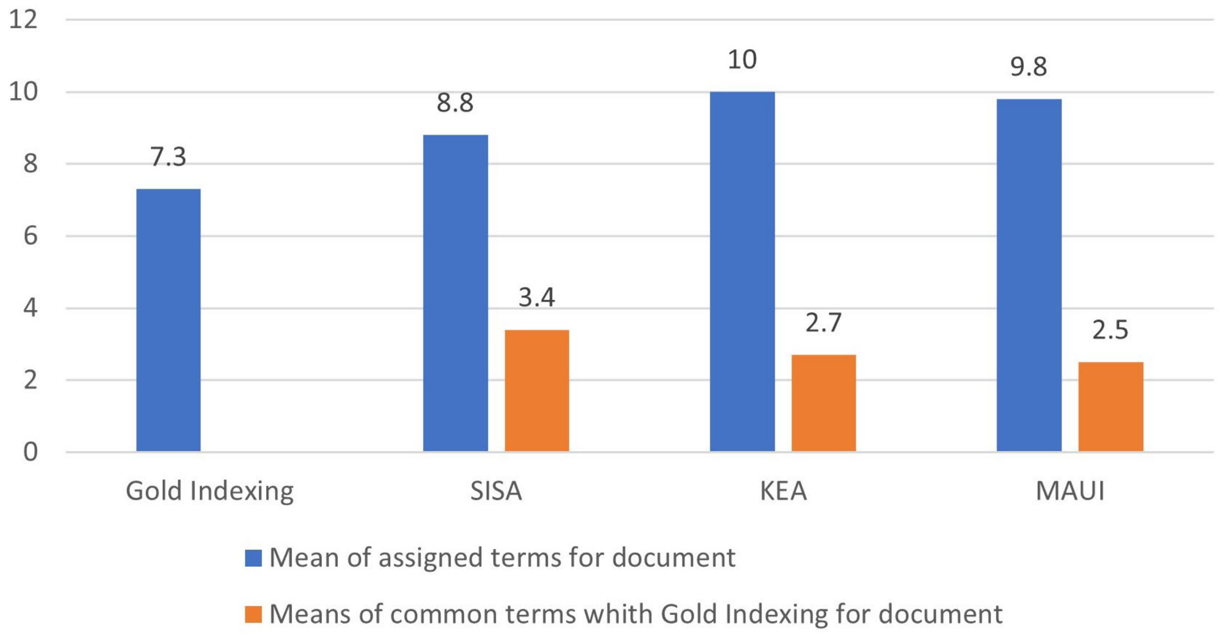


Table II: Measurements used and data for article number 4 in the three systems

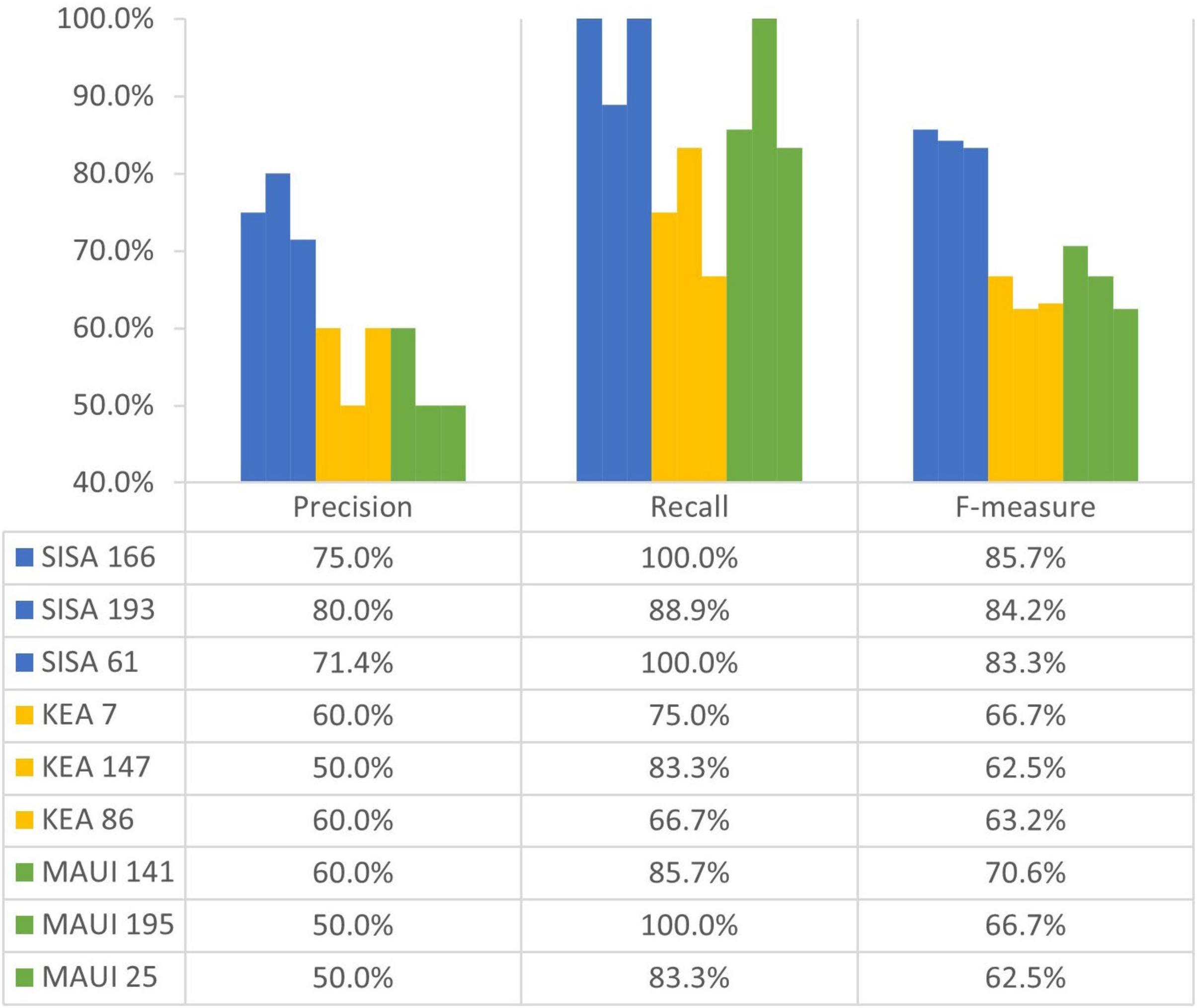
System					Precision (%)	Recall (%)	F-measure (%)
	D	C	A	GI	P = (C/A)	R= (C / GI)	F1=2*(P*R)/(P+R)
SISA	A4	3	11	4	27.27	75.00	40.00
KEA	A4	2	10	4	20.00	50.00	28.57
MAUI	A4	3	10	4	30.00	75.00	42.86

Table II shows that, overall, the values of the metrics obtained by each system for article number 4 are different, with the exception of the similarity observed in the results obtained by SISA and MAUI.

Graph 2 shows the values of the metrics for the three documents where each automatic indexing system has performed best.

It can be observed that the best performance for each system is based on different documents. In addition, SISA has obtained higher values for precision, recall and F-measure. The values obtained by KEA and MAUI for F-measure are similar to each other, but KEA has better values for precision while MAUI has better values for recall.

Graph 2: The three best values for each system for precision, recall and F-measure



	Precision	Recall	F-measure
🟦 SISA 166	75.0%	100.0%	85.7%
🟦 SISA 193	80.0%	88.9%	84.2%
🟦 SISA 61	71.4%	100.0%	83.3%
🟨 KEA 7	60.0%	75.0%	66.7%
🟨 KEA 147	50.0%	83.3%	62.5%
🟨 KEA 86	60.0%	66.7%	63.2%
🟩 MAUI 141	60.0%	85.7%	70.6%
🟩 MAUI 195	50.0%	100.0%	66.7%
🟩 MAUI 25	50.0%	83.3%	62.5%

On the other hand, the three worst non-zero results for the F-measure range between 5.9% and 11.8% for KEA, 7.4% and 10.5% for SISA, and 8.3% and 9.5% for MAUI. Thus, SISA, KEA and MAUI have similar values for precision, recall and F-measure in the worst cases.

Table III shows the statistics for the metric values achieved by the automatic indexing systems based on the test set.

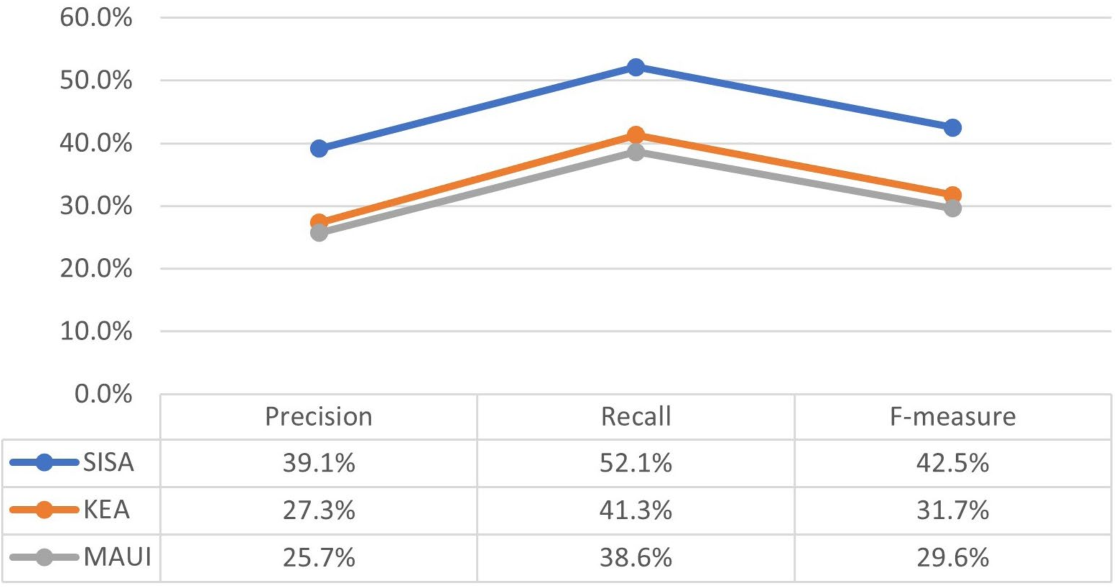
According to unpaired t test results with conventional criteria, the difference between the means of the values of the metrics for SISA and those of each of the other systems is considered extremely statistically significant. However, according to unpaired t test results with conventional criteria, the difference between the means of the values of the metrics for KEA and MAUI is not statistically significant.

Graph 3 illustrates the similar mean performance of KEA and MAUI, and the superior mean performance of SISA.

Table III: Statistics for SISA, KEA and MAUI performance on test set

		Assigned terms	Gold indexing	Mean of common terms	Precision	Recall	F-measure
SISA	Minimum	3	2	0	0%	0%	0%
	Mean	8.9	7.3	3.4	39.1%	52.1%	42.5%
	Deviation	2.7	3.4	1.7	17.9%	27.7%	18.9%
	Maximum	18	24	8	100%	100%	85.7%
KEA	Minimum	10	2	0	0%	0%	0%
	Mean	10	7.3	2.7	27.3%	41.3%	31.7%
	Deviation	0	3.5	1.3	12.9%	22.1%	14.6%
	Maximum	10	24	6	60.0%	100%	66.7%
MAUI	Minimum	6	2	0	0%	0%	0%
	Mean	9.8	7.3	2.5	25.7%	38.6%	29.6%
	Deviation	0.6	3.4	1.3	13.3%	22.4%	14.9%
	Maximum	10	24	6	60.0%	100%	70.6%

Graph 3: Summary table with the resulting data



	Precision	Recall	F-measure
SISA	39.1%	52.1%	42.5%
KEA	27.3%	41.3%	31.7%
MAUI	25.7%	38.6%	29.6%

Table IV shows the number of documents for which the performance of the automatic indexing systems falls in the intervals of F-measure values.

Table IV: F-measure values of the two hundred indexed documents

	F-measure 0%	F-measure 1-15%	F-measure 16-49%	F-measure ≥50%
SISA	6	11	111	**72**
KEA	8	26	**145**	21
MAUI	**10**	**34**	135	21

We can see in Table IV that the systems have the majority of the F-measure values falling in the interval of 16 to 49% for F-measure. SISA has roughly three times the number of Fmeasure values equal to or higher than 50% compared to KEA and MAUI. KEA and MAUI have similar numbers of documents in each interval of F-measure values. In addition, the number of documents with an F-measure of zero is as follows: SISA produced six, KEA eight, and MAUI ten.

According to Medelyan (2009), the similarity between the values achieved by KEA and the data for MAUI is explained by the fact that MAUI was developed from the foundations and structure of KEA. However, we can observe that besides their similar performance, there is a performance variation for each document in these systems, so that they capture both common and different aspects of the statistical nature of indexing terms.

SISA was conceived with the aim of mimicking a human indexer by directing the focus towards places in the text where it can identify significant terms. ISO standard 5963-1985 on human indexing of documents states that "important parts of the text need to be considered carefully, and particular attention should be paid to the following: a) the title, b) the abstracts, if provided; c) the list of contents; d) the introduction, the opening phrases of chapters and paragraphs, and the conclusion; e) illustration, diagrams, tables and their captions". This recommendation by the standard derives from the experience of human indexers, hence SISA's imitation of the behaviour of human indexers is perhaps the reason why SISA has achieved better results than KEA and MAUI, which do not base their processing on the structural positions of terms in documents.

3.4 Consistency

Indexing consistency seeks to determine the similarity between sets of indexing terms from analysis of the same document. Therefore, there are a number of possible combinations for comparison: between human indexing, between automatic indexing, and between human indexing and automatic indexing.

Hooper's measure has been extensively used to calculate indexing consistency. In the formula, C is the number of terms assigned by SISA, KEA and MAUI that match those assigned by golden indexing, A is the number of terms assigned by SISA, KEA or MAUI, and GI is the number of golden indexing terms for a given document. Table V summarizes the data achieved in our experiment.

Table V: Hooper's measure

		Hooper H = C / (A+GI-C)
SISA	Minimum	0%
	Mean	**28.8%**
	Deviation	15.8%
	Maximum	75.0%
KEA	Minimum	0%
	Mean	19.7%
	Deviation	10.6%
	Maximum	50.0%
MAUI	Minimum	0%
	Mean	18.3%
	Deviation	10.7%
	Maximum	54.5%

C= Number of commons terms assigned by SISA, KEA and MAUI in the relation to gold indexing; A= Number of terms assigned by SISA, KEA or MAUI; GI= Number of gold indexing terms for a given document.

Graph 4 shows the three documents with the highest consistency. SISA's levels of consistency are higher than those of KEA and MAUI. KEA and MAUI have similar values of consistency for the best cases, but for different documents.

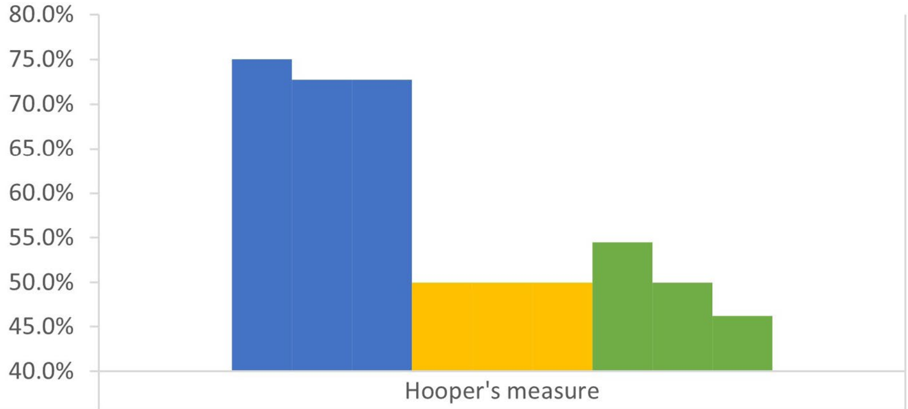
The three worst results achieved a range from 3.8% to 5.8% for SISA, 3% to 6.3% for KEA and 4.3% to 5% for MAUI. Again, the systems generate very similar values for the three worst cases.

3.5 Indexing analysis

In this section we analyze the three best performance results and the three worst non-zero results achieved by the three systems.

In Table VI we show the indexing that achieved the highest F-measure.

Graph 4: Hooper's best three values



	Hooper's measure
🟦 SISA 166	75.0%
🟦 SISA 62	72.7%
🟦 SISA 193	72.7%
🟨 KEA 7	50.0%
🟨 KEA 167	50.0%
🟨 KEA 178	50.0%
🟩 MAUI 141	54.5%
🟩 MAUI 195	50.0%
🟩 MAUI 193	46.2%

Table VI: Indexing with the best F-measures

SISA (Article 166). F-measure = 85.7%	
Gold indexing	**SISA indexing**
1. **Cadena de valor** 2. Desintermediación 3. **Digitalización** 4. **España** 5. **Industria editorial** 6. **Libros electrónicos**	1. **Cadena de valor** 2. Desinformación 3. **Digitalización** 4. **España** 5. Estudio 6. Impacto 7. **Industria editorial** 8. **Libros electrónicos**
KEA (Article 7). F-measure = 66.6%	
Gold indexing	**KEA indexing**
1. **Accesibilidad web** 2. Andalucía 3. **Discapacidad** 4. **Diseño web** 5. Sitios web 6. **Universidades** 7. **W3C** 8. **WCAG 2.0**	1. Accesibilidad 2. **Accesibilidad web** 3. **Discapacidad** 4. **Diseño Web** 5. Personas con discapacidad 6. Portales 7. **Universidad** 8. **W3C** 9. **WCAG 2.0** 10. World Wide Web
MAUI (Article 141). F-measure = 70.5%	
Gold indexing	**MAUI indexing**
1. **Bibliotecas** 2. **Bibliotecas escolares** 3. **Blogs** 4. **Educación infantil** 5. España 6. **Extremadura** 7. **Indicadores de calidad**	1. **Bibliotecas** 2. **Bibliotecas escolares** 3. **Blogs** 4. Centros de enseñanza 5. Centros educativos 6. **Educación infantil** 7. **Extremadura** 8. **Indicadores de calidad** 9. Modelo 10. Modelos

In Table VI, we can see in bold type the terms that golden indexing and the automatic indexing had in common with each other for the best cases. The number of terms they had in common is five or six terms. The terms proposed by automatic indexing which were not common to both the automatic and manual systems are however linked in some way to the golden indexing terms for each document. In addition, the number of terms in golden indexing and the number proposed by the systems is similar. In Annex, Table A shows the three best results with the same patterns as those indicated above.

Table B in Annex shows the worst non-zero results for SISA, KEA and MAUI. The number of terms they had in common is one or two terms. The terms proposed by automatic indexing which were not common to both the automatic and manual systems are linked in some way to the golden indexing terms for each document. In addition, a greater difference between the number of terms in golden indexing and the number of those proposed by the systems can be observed for some documents. In Annex, Table A shows the three worst non-zero results with the same patterns as those indicated here.

In Annex 1, Table A shows the three best and three worst non-zero results for SISA, KEA and MAUI. In the case of SISA, there is a considerable disparity between the number of terms assigned by SISA and the number of terms assigned by the database's human indexers. In the case of Article 91, the human indexers assigned almost three times as many terms as SISA while for Article 94 they assigned twice as many. For Article 87, SISA assigned almost twice as many terms as those assigned by human indexing. This large disparity between SISA and human indexing also occurred for other articles that achieved a low F-measure of between 10-16% (Articles 1, 97, and 168, for example), although in other articles with a very similar number of assigned terms (Articles 10, 36 and 138) the F-measure was equally low.

If the indexing by a given tool gives an F-measure that is similar to the examples shown in Table VI, we consider that use of some automatic indexing software or semi-automatic could be of great help in order to increase document processing capacity in real working environments, or even by enabling the automatic indexing of documents that are not manually indexed.

4. CONCLUSIONS AND FURTHER WORK

This study proposed an evaluation-comparison between three automatic indexing systems, analyzing, on the one hand, the latest web version of SISA (which uses a rule-based algorithm focusing on the position occupied by the terms in the documents) and, on the other hand, KEA and MAUI (two indexing systems that produce indexing terms by means of a machine learning model, after a training process with 30 documents) and their respective indexing.

The data achieved by SISA with a total F-measure of ten points above KEA and MAUI, more than three times as many documents with an F-measure ≥50% and an average number of terms assigned per document of 8.8, indicate that the automatic indexing by SISA is more similar to the human indexing by the InDICES database professionals than the automatic indexing by KEA and MAUI. Thus, an algorithm that focuses on different parts of the texts (titles, abstracts, author keywords, headings, first paragraphs of headings, table titles, graph titles, conclusions and references) and exploits the structure of documents in XML format has outperformed algorithms based on machine learning and statistical features of terms.

By exploiting markup tags and being capable of handling XML documents and documents generated from JATS, SISA is in line with the current trend of creating publications that allow further automatic processing and reuse. It is therefore also worth mentioning that JATS has been the basis for creating two extensions: the Book Interchange Tag Suite (BITS), which is an XML model to describe the structural and semantic content of books published by scientific, technical and medical publishers, and the STS (Standards Tag Suite), ANSI/NISO Z39.102-2017 and the ISOSTS (ISO Standards Tag Set) systems which provide a common tagging format that developers, publishers and distributors of standards can use to publish and exchange contents. Therefore, SISA seems to be moving in the right direction if these forms of publication are extended.

With regard to future work, it would be useful to replicate the experiments presented here with texts from other disciplines, other professional human indexing and other controlled vocabularies in order to verify whether the higher performance level achieved by the methodology implemented in SISA still outperforms the KEA and MAUI machine learning algorithms. It would perhaps also be interesting to work jointly with the indexers of the InDICES database to carry out a detailed analysis of the indexing of the 200 articles performed by SISA, KEA and MAUI so as to determine, for example, which types of articles obtained the greatest similarity with human indexing or to establish what types of errors are made by the automatic systems, among other aspects. Moreover, during the execution of this work, several improvements have been identified that could improve SISA's performance. One of them could be to limit the maximum number of terms that can be assigned per document, which would perhaps allow higher F-measure indexes to be achieved. In this regard, it should be noted that MAUI has limited the maximum number of indexing terms it can assign to 10 and KEA systematically establishes 10 terms for each document, whereas SISA currently has neither a minimum nor a maximum number, and thus this study has observed one document with 18 terms assigned by SISA, several with three and many others with 12, 13 and 14 terms. Another improvement that is already being implemented is a new module to generate indexing rules automatically. In SISA, indexing terms are derived by applying a set of rules established by the user according to their experience and knowledge of the system. With this improvement, the aim is to eliminate user intervention in favour of a data-driven automatic process that generates rules from a collection of training documents with their respective golden indexing, so that it is possible to know which rules produce F-measure values above certain thresholds. This data-driven automatic process would permit greater adaptability by SISA to the characteristics of the texts and subject areas of each test set. Finally, it would also be interesting to find out if the automatically generated rules manage to improve on the results of the manual rules established and used for SISA in the execution of the experiments presented here.

5. ACKNOWLEDGMENTS

The authors would like to thank Teresa Abejón Peña for providing us with the controlled vocabulary used in the InDICES database to carry out the experiments. We would also like to extend our thanks to the Consejo Superior de Investigaciones Científicas, which is responsible for this database. Lastly, we would like to thank two anonymous reviewers for their comments, which have helped improve the understanding of the manuscript.

AGRADECIMIENTOS

Los autores agradecen a Teresa Abejón Peña por facilitarnos el vocabulario controlado utilizado en la base de datos InDICES para la realización de los experimentos. También extendemos el agradecimiento al Consejo Superior de Investigaciones Científicas como responsable de esta base de datos.Por último, agradecemos a los dos revisores anónimos sus comentarios, que han ayudado a mejorar la comprensión del manuscrito.

6. REFERENCES

Aronson, A.R., Bodenreider, O., Chang, H., Florence, H., Humphrey, S.M., Mork, J. G., Stuart, J.N., Rindflesch, T. C., & Wilbur, W. J. (2000). The NLM Indexing Initiative. In J. Marc Overhage (ed.), *Proceedings of the AMIA Annual Symposium*, 17-21.

Akhtar, N., Javed, H., & Ahmad, T. (2017). Searching related Scientific Articles Using Formal Concept Analysis. In *International Conference on Energy, Communication, Data Analytics and Soft Computing* (ICECDS), 2158-2163. DOI: https://doi.org/10.1109/ICECDS.2017.8389834

Amat, N. (1989). *Documentación y nuevas tecnologías de la información*. Pirámide.

Al-Zoghby, A. (2018). *A New Semantic Distance Measure for the VSM-Based Information Retrieval Systems*. In *Intelligent Natural Language Processing: Trends and Application,* 740: 229-250. https://doi.org/10.1007/978-3-319-67056-0_12

Aquino, G., & Lanzarini, L. (2015). Keyword Identification in Spanish Documents using Neural Networks. *Journal of Computer Science and Technology*, 15, 55-60.

Bandim, M. A. S., & Corrêa, R. F. (2019). Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação. *Transinformação*, 31, 1-12. https://doi.org/10.1590/2318-0889201931e180004

Chebil, Wiem, Soualmia, L., Dahamna, B., & Srmoni, S. (2012). Indexation automatique de documents ensanté: évaluation et analyse de sources d'erreurs. *IRBM*. 33, 316-329. DOI: https://doi.org/10.1016/j.irbm.2012.10.002

Cleverdon, C.W. (1962). *Aslib Cranfield Research Project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield.

Duwairi, R., & Hedaya, M. (2016). Automatic keyphrase extraction for Arabic news documents based on KEA system. *Journal of Intelligent and Fuzzy Systems*, 30(4), 2101-2110.

El-Haj, M., Balkan, L., Barbalet, S., Bell, L., & Shepherdson, J. (2013). An Experiment in Automatic Indexing Using the HASSET Thesaurus. In *5th Computer Science and Electronic Engineering Conference (CEEC),* 13-18. DOI: https://doi.org/10.1109/CEEC.2013.6659437

Evans, D. A. (1990). Concept Management in Text via Natural-Language Processing: the CLARIT Approach. In *Working Notes of the 1990 AAAI Symposium on "Text-Based Intelligent Systems'9*, Stanford University, March, 27-29, 93-95.

Evans, D.A., Hersh W.R., Monarch, I., Lefferts, R. G., & Handerson, S. K. (1991a). Automatic Indexing of abstracts via Natural-Language Processing Using a Simple Thesaurus. *Medical Decision Making*, 11(4), 108-115.

Evans, D.A., Handerson, S. K., Lefferts, R. G., & Monarch, I. (1991b). A Summary of the CLARIT Project. November 1991, Report No. CMU-LCL-91-2. DOI: https://doi.org/10.1184/R1/6490799.v1

Farrow, J. (1994). Indexing as a cognitive process. In Kent, A., Lancour, H. and Daily, J.E. (eds). *Encyclopedia of Library and Information Science*, 53, 155-171.

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm*, Sweden, 668–673. San Francisco, CA: Morgan Kaufmann Publishers.

García Gutiérrez, A. (1984). *Lingüística documental*. Barcelona: Mitre.

Gil-Leiva, I. (2008). *Manual de indización. Teoría y práctica*. Trea.

Gil-Leiva, I. (2017a). SISA: Automatic Indexing System for Scientific Articles. Experiments with Location Heuristics Rules versus TF-IDF Rules. *Knowledge Organization*, 44(3), 139-162.

Gil-Leiva, I. (2017b). La indización de artículos científicos con el sistema de indización automática SISA comparada con la indización en las Bases de datos Agricola, WoS y SCOPUS. In *Third Spanish-Portuguese ISKO Conference, Portugal, Thirteenth ISKO Conference, Spain*, University of Coimbra, 23 and 24 November, 510-524.

Gopan, E., Rajesh, S. Gr, V., Akhil, R. R., & Thushara, M. (2020). Comparative Study on Different Approaches in Keyword Extraction. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 70-74. DOI: https://doi.org/10.1109/iccmc48092.2020.iccmc-00013

Gupta, Y., Saini, A., & Saxena, A. (2015). A new fuzzy logic based ranking function for efficient Information Retrieval system. *Expert Systems with Applications*, 42(3), 42, 1223-1234.

Hersh W. R., & Greenes R. (1990). SAPHIRE: An information Retrieval Environment Featuring Conceptmatching, Automatic Indexing, and Probabilistic Retrieval. *Computers and Biomedical Research*, 123, 410-425.

Hersh W. R., Hickam D. H., Haynes, R. B., & McKibbon, K. A. (1991). Evaluation of SAPHIRE: an Automated Approach to Indexing and Retrieving Medical Literature. In *ProceedingsSymposium on Computer Applications in Medical Care*, 808-812.

Hooper, R.S. (1965). *Indexer consistency tests: origin, measurement, results, and utilization*. IBM Corporation, (TR95-56).

Humphrey, S. M., & Miller, N. E. (1987). Knowledge-Based Indexing of the Medical Literature: The Indexing Aid Project. *Journal of the American Society for Information Science*, 38(3), 84-196.

Humphrey, S. M. (1999). Automatic Indexing of Documents from Journal Descriptors: A Preliminary Investigation. *Journal of the American Society for Information Science*, 50(8), 661-674.

Humphrey, S. M., Rogers, W. J., Kilicoglu, H., Demner-Fushman, D., & Rindflesch, T. C. (2006). Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment. *Journal of the American Society for Information Science and Technology*, 57(1), 96-113.

Irfan, R., Khan, S., Qamar, A. M., & Bloodsworth, P. C. (2014). Refining Kea++ Automatic Keyphrase Assignment. *Journal of Information Science*, 40(4), 446-459. DOI: https://doi.org/10.1177/0165551514529054

Irving, H. B. (1997). Computer-assisted Indexing Training and Electronic Text Conversion at NAL. *Knowledge Organization*, 24(1), 4-7.

ISO 5963:1985 : Documentation -- Methods for Examining Documents, Determining their Subjects, and Selecting Indexing Terms. Geneva: ISO.

Karetnyk, D., Karlsson, F., & Smart, G. (1991). Knolewledge-based Indexing of Morpho-Syntactically Analysed Language. *Expert Systems for Information Management*, 4(1), 1-29.

Khan et al. (2011). A Refined Methodology for Automatic Keyphrase Assignment to Digital Documents. *Journal of Digital Information Management*, 9(2), 55-63.

Kim, S. N., Medelyan, O., Kan, M., & Baldwin, T. (2013) Automatic Keyphrase Extraction from Scientific Articles. *Language Resources and Evaluation*, 47, 723–742. DOI: https://doi.org/10.1007/s10579-012-9210-3

Klingbiel, P. H. (1973). A Technique for Machine-Aided Indexing. *Information Storage and Retrieval*, 9(9), 477-494. DOI: https://doi.org/10.1016/0020-0271(73)90034-X

Krapivin, M., Marchese, M., Yadrantsau, A, & Liang, Y. (2008). Unsupervised Key-Phrases Extraction from Scientific Papers using Domain and Linguistic Knowledge. In *International Conference on Digital Information Management*, 105-112.

Lima, V. M. A., & Boccato, V. R. C. (2009). O desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP nos processos de indexação manual, automática e semi-automática. *Perspectivas em Ciência da I*nformação, 1, 131-151.

Lin, N., Kudinov, V.A., Zaw, H.M., & Naing, S. (2020). Query Expansion for Myanmar Information Retrieval Used by WordNet. In*2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering* (EIConRus), 395-399.

Medelyan, O. (2005). *Automatic keyphrase indexing with a domain-specific thesaurus*. Master's thesis, Albert-Ludwigs University.

Medelyan, O. (2009). Human-competitive automatic topic indexing. PhD Thesis. University of Waikato, New Zealand. Available at: https://cds.cern.ch/record/1198029/files/Thesis-2009-Medelyan.pdf [Consulted: 05/05/2021].

Mork, J. G., Aronson, A., & Demner-Fushman, D. (2017). 12 Years on – Is the NLM Medical Text Indexer Still Useful and Relevant?. *Journal of Biomedical Semantics***,** 8. DOI: https://doi.org/10.1186/s13326-017-0113-5

Mynarz, J., & Škuta, C. (2010). Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository. In *Twelfth International Conference on Grey Literature*, Prague, December. Available at: http://www.nusl.cz/ntk/nusl-42005 [Date consulted: 24/03/2021].

Névéol, A., Mary, V., Gaudinat, A., Boyer, C., Rogozan, A., & Darmoni, S. J. (2005). A Benchmark Evaluation of the French MeSH Indexers. *Lecture Notes in Computer Science*, 251–255. DOI: https://doi.org/10.1007/11527770_37

Rae, A., Pritchard, D., Mork, J. G., & Emner-Fushman, D. (2021). Automatic MeSH Indexing: Revisiting the Subheading Attachment Problem. In *Annual Symposium proceedings. AMIA Symposium*, 2020, 1031-1040.

Rolling, L. N. 1981. Indexing Consistency, Quality snd Efficiency. *Information Processing and Management,* 17, 69-76.

Salisbury, L., & Smith, J. J. (2014). Building the AgNIC Resource Database Using Semi-Automatic Indexing of Material. *Journal of Agricultural & Food Information*, 15 (3), 159-176. DOI: https://doi.org/10.1080/10496505.2014.919805

Salton, G. (1989). The SMART system 1961-1976: Experiments in Dynamic Document Processing. *Encyclopedia of Library and Information Science*, 28, 1-28.

Salton, G. (1991). The Smart Document Retrieval Project. In *Proceeding SIGIR '91 Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 356-58.

Scholastica survey: The State Of Journal Production And Access 2020. Available at: https://lp.scholasticahq.com/journal-production-access-survey/[Date consulted: 8/10/2021].

Seiler, M., Hübner, P., & Paech, B. (2019). Comparing Traceability through Information Retrieval, Commits, Interaction Logs, and Tags. In *2019 IEEE/ACM 10th International Symposium on Software and Systems Traceability (SST)*, 21-28.

Shams, R., & Mercer, R. E. (2012a). Investigating Keyphrase Indexing with Text Denoising. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '12*. DOI: https://doi.org/10.1145/2232817.2232866

Shams, R., & Mercer, R.E. (2012b). Improving Supervised Keyphrase Indexer Classification of Keyphrases with text Denoising. *Lecture Notes in Computer Science*, 77-86.

Silva, S. R. de B., & Corrêa, R. F. (2020). Sistemas de Indexação automática por atribuição: uma análise comparativa. *Encontros Bibli: Revista Eletrônica De Biblioteconomia E Ciência Da Informação*, *25*, 1-25. DOI: https://doi.org/10.5007/1518-2924.2020.e70740

Silva, S. R. de B., & Corrêa, R. F., Gil-Leiva, I. (2020). Avaliação direta e conjunta de Sistemas de Indexação Automática por Atribuição. *Informação & Sociedade-Estudos*, 30, 1-27. http://dx.doi.org/10.22478/ufpb.1809-4783.2020v30n4.57259

Silvester, J. P., Genuardi, M. T., & Klingbiel, P. H. (1994). Machine-Aided Indexing at NASA. *Information Processing & Management* 30 (5), 631-645.

Sinkkilä, R., Suominen, O., & Hyvönen, E. (2011). Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages. *Proceedings The Semantic Web: Research and Applications : 8th Extended Semantic Web Conference*, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 215–229. DOI: https://doi.org/10.1007/978-3-642-21034-1_15

Souza-Rocha, R., & Gil-Leiva, I. (2016). Automatic Indexing of Scientific Texts: A Methodological Comparison. In Chaves Guimarães, J. A., Oliveira Milani, S., Dodebei, V., *Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society: Proceedings of the Fourteenth International ISKO Conference* 27-29 September 2016, 243-250. Rio de Janeiro, Brazil. Würzburg: Ergon Verlag.

Suominen, O. (2019). Annif: DIY Automated Subject Indexing using Multiple Algorithms. *LIBER Quarterly*, 29 (1), 1-25. DOI: http://doi.org/10.18352/lq.10285

Wang, D.X., Gao, X., & Andreae, P. (2015). DIKEA: Exploiting Wikipedia for keyphrase extraction. *Web Intelligence*, 13 (3), 153-165.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, 254-255, 243-250 https://doi.org/10.1145/313238.313437

7. ANNEX

Table A: The three best results and three non-zero worst results of SISA, KEA and MAUI

	Article	Assigned terms	Terms of Gold indexing	Terms commons	Precision	Recall	F-measure
	166	8	6	5	75%	100%	85%
S	193	10	9	8	80%	88%	84%
I	61	7	5	5	71%	100%	83%
S	91	7	20	2	14%	5%	7.4%
A	87	13	7	1	7%	14%	10.0%
	94	6	12	1	16%	8%	11.1%
	7	10	8	6	60%	75%	66%
K	147	10	6	5	50%	83%	62%
E	86	10	9	6	60%	66%	63%
A	133	10	8	1	10%	12%	11.1%
	17	10	7	1	10%	14%	11.7%
	18	10	6	1	10%	16%	12.5%
	141	10	7	6	60%	85%	70%
M	195	10	5	5	50%	100%	66%
A	193	10	9	6	60%	66%	63.1%
U	131	10	14	1	10%	7%	8.3%
I	119	10	11	1	10%	9%	9.5%
	16	10	9	1	10%	11%	10.5%

Table B: Indexing terms of the non-zero worst results of SISA, KEA and MAUI

SISA (Document 91). F-measure = 7.4%	
Gold indexing	**SISA indexing**
1. Acceso a la información 2. Análisis bibliográfico 3. Archivos abiertos 4. Autenticación 5. Ciudadanos 6. Datos abiertos vinculados 7. Demanda de información 8. Documentación 9. Documentos 10. Estudios de casos 11. Gestión de la información 12. Información 13. **Información pública** 14. Instituciones públicas 15. Internet 16. Publicaciones oficiales 17. Reciclaje 18. Reutilización 19. **Sector público** 20. Tecnologías de la información y la comunicación (TIC)	1. Acceso 2. Datos 3. **Información pública** 4. Reutilización de información 5. **Sector público** 6. Tic 7. Usuarios
KEA (Document 133). F-measure = 11.1%	
Gold indexing	**KEA indexing**
1. Comunicación 2. Deontología 3. Editores 4. Educación 5. España 6. Ética 7. Psicología 8. **Revistas científicas**	1. Aspectos éticos 2. Ciencias Sociales 3. Colaboración científica 4. CSIC 5. Humanidades 6. Psicología de la Educación 7. Revistas 8. **Revistas científicas** 9. Trabajo de investigación 10. Universidad
MAUI (Document 131). F-measure = 8.3%	
Gold indexing	**MAUI indexing**
1. Carrera profesional 2. CERN 3. Científicos 4. Colaboración científica 5. Estudios de casos 6. Experimentación científica 7. Ginebra 8. Historia de la ciencia 9. Investigadores 10. Organización de la investigación 11. Proyecto Atlas 12. Relaciones laborales 13. Suiza 14. **Transmisión de conocimientos**	1. Atlas 2. Cooperación 3. Desarrollo profesional 4. Experimento 5. Física de partículas 6. Miembros 7. Organización social 8. Toma de decisiones 9. Trabajadores 10. **Transmisión de conocimientos**