# Evaluación de Herramientas de Extracción Automática de Conceptos Dentro de un Ambiente de Biblioteca Digital

Rocío Abascal [1] Béatrice Rumpler[1]

**Resumen**

El rápido avance de la tecnología ha originado la proliferación de fuentes de información digital. Esta evolución informática ha provocado la creación de bibliotecas digitales que han ido convirtiendose poco a poco en un gran pilar para la difusión del conocimiento. Sin embargo, la información contenida en las bibliotecas digitales aún no está descrita totalmente y su explotación es aún insuficiente. Recientemente, se ha comprobado que la descripción de la información usando *"metadatos"* puede ser primordial para el mejoramiento de la consulta de la información dentro de una biblioteca digital. Nuestro enfoque está basado en la creación e introducción de nuevos *"metadatos"* capaces de describir, en nuestro caso, las tesis doctorales de una biblioteca digital. Estos *"metadatos"* corresponden a los conceptos más importantes de cada una de las tesis. Actualmente, la identificación manual de conceptos es un largo proceso llevado a cabo por un especialista del área. Por lo tanto, es importante hacer uso de herramientas capaces de extraer automáticamente conceptos. En este artículo analizamos cuatro herramientas de PLN (Procesamiento del Lenguaje Natural) capaces de extraer automáticamente los conceptos claves de un corpus. Estas herramientas son: (1) TerminologyExtractor de Chamblon Systems Inc., (2) Xerox Terminology Suite de Xerox, (3) Nomino de Nomino Technologies y (4) Copernic Summarizer de NRC. Este artículo presenta también un prototipo de herramienta de anotación desarrollado para insertar de manera automática conceptos a las tesis digitales.

**Palabras claves:** *Biblioteca digital, metadatos, Procesamiento del Lenguaje Natural, extracción de información, anotación, búsqueda de información.*

**Abstract**

The fast advance of the technology has originated the proliferation of digital sources of information. This computer evolution has caused the creation of digital libraries that have become a big pillar for the diffusion of knowledge. However, the information contained in the digital libraries is not totally described and its exploitation is still insufficient. Recently, it has been proven that describing the information by using *"metadata"* can be fundamental for the improvement of the research of the information within a digital library. Our approach is based on the creation and the introduction of new *"metadata"* able to describe, in our case, the PhD theses of the digital library. These *"metadata"* correspond to the most important concepts of each one of the theses contained in the digital library. At the moment, manual identification of concepts is a long process that is carried out by a specialist of the area. Therefore, we considered the use of tools to be able to automatically extract concepts. In this article we analyze four tools of NLP (Natural Language Processing) able to automatically extract the key concepts of a corpus. These tools are: (1) TerminologyExtractor of Chamblon Systems Inc., (2) Xerox Terminology Suite of Xerox, (3) Nomino of Nomino Technologies and (4) Copernic Summarizer of NRC. This paper also presents a prototype developed to automatically insert concepts into digital theses.

**Keywords:** *Digital library, metadata, Natural Language Processing, information extraction,annotation, information research.*

[1] INSA de Lyon – LIRIS, 7 Avenue J. Capelle Bät 502 – Blaise Pascal F69621 Villeurbanne cedex FRANCE. {Rocio.Abascal, Béatrice.Rumpler}@insa-lyon.fr

# 1 Introducción

A pesar de que los sistemas de recuperación de información son cada día más sofisticados y precisos, las necesidades del usuario aún no están completamente cubiertas. Los usuarios todavía deben juzgar sobre la pertinencia de los documentos obtenidos al realizar una búsqueda de información. Algunos de los criterios utilizados todavía por los motores de búsqueda para evaluar si la información es pertinente o no, se basan solamente en el número de veces en que las palabras clave aparecen en los documentos. Sin embargo, los usuarios deben leer grandes fragmentos del documento o en el peor caso el documento por completo para saber si es pertinente o no a sus necesidades. Con el fin de atacar este problema, nuestro enfoque está basado en la descripción conceptual de los documentos ya que consideramos que es fundamental para poder poner al alcance la información pertinente a los usuarios.

Nuestro estudio se realiza dentro del proyecto llamado CITHER[2] (Consulta integral de tesis en red). CITHER, desde 1997, lleva a cabo la difusión, vía Internet, de las tesis doctorales del INSA de Lyon, Francia. La difusión de las tesis se realiza utilizando el formato PDF (Portable Document Format). La descripción de ciertos datos correspondientes a la tesis se realiza actualmente a través del uso del formato Dublin Core. A partir de estos metadatos el sistema se apoya para realizar búsquedas de información. Actualmente, el usuario puede efectuar una consulta utilizando palabras correspondientes al título, el nombre del autor, la fecha o el año de edición. Sin embargo, la mayor parte del tiempo el usuario no conoce estos datos y por lo tanto no obtiene resultados pertinentes al efectuar una consulta. CITHER permite obtener la tesis completa durante una consulta pero cuando la tesis no es pertinente sería necesario obtener previamente fragmentos que en el caso de que el usuario considerara pertinentes pudiera desplegar en forma completa. Es en éste enfoque en el que trabajamos actualmente: caracterizar las tesis por medio de nuevos metadatos [2] y apoyar la búsqueda de información pertinente usando una ontología del área [1] para ofrecerle al usuario los fragmentos pertinentes de cada tesis.

Con el objetivo de caracterizar mejor el contenido de las tesis, hemos decidido usar nuevos *"metadatos"* basados en los conceptos que describen cada una de las tesis. La extracción manual de los conceptos que caracterizan un documento es una tarea larga y complicada ya que se necesita tener conocimiento del área de especialidad. Cuando se tienen pocos documentos a evaluar tal vez la extracción manual sea una tarea factible, sin embargo, al evaluar grandes cantidades de documentos se necesita que el proceso se lleve a cabo de manera automática o semiautomática. En este caso, hablamos de un proceso semiautomático ya que es el experto el que evaluará la pertinencia de los conceptos extraídos por las herramientas.

Para poder seleccionar una herramienta de PLN (Procesamiento de Lenguaje Natural) capaz de extraer automáticamente los conceptos de un corpus dado como entrada, hemos decidido buscar herramientas que puedan satisfacer nuestras necesidades. En este artículo, presentamos una evaluación empírica de cuatro herramientas que extraen automáticamente conceptos. Estas herramientas son: (1) TerminologyExtractor de Chamblon Systems Inc., (2) Xerox Terminology Suite de Xerox, (3) Nomino de Nomino Technologies y (4) Copernic Summarizer de NRC. La evaluación ha sido realizada mediante la comparación del grado de similitud entre los conceptos extraídos por cada una de las herramientas y una lista de referencia que contiene conceptos extraídos manualmente por un experto del área.

[2] http://docinsa.insa-lyon.fr/docinsa/these/

La estructura del artículo es la siguiente: en la Sección 2 presentamos un resumen de trabajos relacionados con el área de extracción automática de términos. En la Sección 3 presentamos las principales características de las herramientas evaluadas. El método utilizado para evaluar las diferentes herramientas lo presentamos en la Sección 4. La evaluación y los resultados de nuestro estudio son presentados en la Sección 5. En la Sección 6, presentamos un prototipo para una herramienta de anotación capaz de insertar los conceptos extraídos por las herramientas a cada una de las tesis de la biblioteca digital. Los conceptos extraídos corresponden a aquellos encontrados dentro de los capítulos más importantes de la tesis. Para esto, explicaremos rápidamente el análisis semántico que realizamos con el fin de encontrar una organización semántica dentro de las tesis y basarnos en esta a la hora de realizar la extracción de conceptos. Las conclusiones y el trabajo futuro se presentan al final del artículo.

## 2 Trabajo relacionado

Los términos son representaciones lingüísticas de los conceptos de un área en particular. Los términos pueden estar formados por una sola palabra, llamados así *"términos simples"* o también pueden estar formados por dos o más palabras, llamados así *"términos complejos"* [8]. Estos términos cuando son extraídos por una herramienta son llamados también *"candidatos-términos"*, ya que por lo general son términos complejos, es decir, grupos de palabras que pueden en cierto caso revelar cierto conocimiento tratado en un documento dado [12]. De esta manera, las herramientas que realizan una extracción automática de conceptos o términos, pueden ser aplicadas para la construcción de diccionarios especializados, mecanismos de traducción, indexación de libros y bibliotecas digitales, categorización de textos, etc. Además, el uso de conceptos para caracterizar documentos permite que estos conceptos sean usados para la extensión de peticiones de búsqueda y para facilitar al usuario el uso de un vocabulario adecuado. Otras de las muchas aplicaciones que podemos encontrar en el uso de conceptos es la capacidad para visualizar estos conceptos dentro del documento y hacer comparaciones de similitud entre documentos. Entre algunos de los trabajos que se orientan hacia la similitud de documentos basada en la extracción de conceptos podemos citar el trabajo de [9] que presenta un motor de búsqueda llamado Keyphind. Este motor permite clasificar los documentos de la colección utilizando los conceptos extraídos a través de la herramienta Kea [10].

La extracción de *"términos complejos"* está basada en algunas de las siguientes suposiciones:

(1) Los textos especializados contienen muchos términos que son siempre utilizados para referirse a los conocimientos o temas de un área especializada;

(2) Cuando un término es significativo tiene grandes posibilidades de que sea utilizado varias veces dentro del texto;

(3) La gran parte de los términos complejos están formados por más de una palabra, la cual raramente es utilizada sola;

(4) Los términos complejos están formados por un nombre o sustantivo seguido de un adjetivo; por ejemplo: *" biblioteca digital"* o *"inteligencia artificial"*.

Algunos algoritmos y métodos nuevos para la extracción de conceptos son presentados en [3,7,10]. En [6], el autor presenta un análisis de diferentes herramientas de extracción de conceptos a las analizadas en el presente artículo. En [12] el autor presenta una gráfica (basada en los resultados

una tesis de maestría[3] y de otros artículos) de comparación en términos de precisión entre las herramientas: LogiTerm, TermSearch, ATAO, Noun Phrase Search y Nomino.

Algunas técnicas utilizadas para determinar cuándo un término es complejo, están basadas en dos estrategias: (1) las *"estrategias lingüísticas"* y (2) las *"estrategias estadísticas o probabilísticas"*. Las *"estrategias lingüísticas"* utilizan conocimientos del idioma tratado, por ejemplo, basandose en ciertas categorías gramaticales. Las *"estrategias estadísticas o probabilísticas"* utilizan la frecuencia como método para la evaluación y la extracción de términos compuestos, basandose en que cuando un término es pertinente entonces significa que será utilizado varias veces dentro del documento. La herramienta *"Copernic Summarizer"* es evaluada en comparación a la herramienta *"IAI-extractor"* en [8]. Estas dos herramientas son un ejemplo de herramientas que usan la frecuencia para la extracción de términos.

Otro trabajo llevado a cabo en el área de extracción de conceptos incluye a las herramientas de generación automática de resúmenes. Estas herramientas no extraen automáticamente conceptos pero se nos hace interesante nombrarlas ya que algunas de ellas trabajan de manera muy similar a las herramientas que en este artículo presentamos. De esta forma entre las herramientas de generación automática de resúmenes podemos mencionar: TextSummary, HyperGen, WebSumm, DataHammer e IntelliScope, las cuales son evaluadas en términos de *"precisión"* y *"recuperación"* en [11].

# 3 Herramientas de extracción automática de conceptos

En esta sección, describimos las cuatro herramientas evaluadas, las cuales son: (1) Copernic Summarizer NRC, (2) Nomino de Nomino Technologies, (3) TerminologyExtractor de Chamblon Systems Inc., y (4) Xerox Terminology Suite de Xerox. Hemos evaluado únicamente las capacidades de cada una de las herramientas para efectuar la extracción automática de conceptos, generalmente llamada *"extracción o adquisición de términos"*.

## 3.1 Criterios para la selección de las herramientas

Los recientes avances en el área del PLN han fomentado la aparición de una nueva generación de herramientas que van más allá de la recuperación tradicional de la información, permitiendo a un sistema entender documentos de forma similar a como lo hacemos las personas, y por tanto posibilitando la extracción de conceptos semánticos útiles. Dentro de estas herramientas encontramos varios tipos de acuerdo a su funcionalidad: (1) análisis sintáctico, (2) extracción de conceptos, (3) etiquetado, (4) manejo del léxico, etc. En nuestro caso, para la selección de las herramientas buscamos únicamente aquellas que llevarán a cabo la extracción de conceptos.

Algunos de los criterios que utilizamos para dicha selección pueden resumirse en:

- *Tratamiento del idioma francés*: muchas de las herramientas que encontramos realizan la extracción automática de conceptos pero muy pocas están destinadas para analizar el idioma francés. Este criterio es de suma importancia ya que nuestro trabajo se realiza dentro de una biblioteca digital de tesis doctorales en francés.

[3] Love, S. Benchmarking the performance of two automated term-extraction Systems, Tesis de maestría, Montreal, Universidad de Montreal, 2000.

- *Facilidad de instalación*: todas las herramientas evaluadas se encuentran en linea y cuentan con una versión de evaluación.

- *Facilidad de uso*: nuestra evaluación tenía que llevarse a cabo en un corto plazo de tiempo por lo cual necesitábamos que todas las herramientas contarán con ejemplos y tutoriales que nos permitieran rápidamente dominar la herramienta.

- *Facilidad para contactar a los autores*: la gran mayoría de las herramientas analizadas, como Nomino, nos permitieron contar con la ayuda de los autores quienes nos mandaron más información al respecto y nos dieron una clave para usar Nomino ilimitadamente.

## 3.2 Copernic Summarizer

El algoritmo de extracción del NRC (National Research Council) es usado por Copernic Summarizer [5] para crear una lista de conceptos a partir de un documento dado. El principal objetivo de Copernic Summarizer es la creación automática de resúmenes [15], sin embargo, lo hemos evaluado sólo por su capacidad para efectuar también la extracción de conceptos. Esta capacidad permite extraer conceptos que están formados por más de una palabra. Al mismo tiempo que Copernic Summarizer crea un resumen para un documento, también produce una lista de conceptos y permite destacarlos dentro del resumen creado. Esta última capacidad, nos ha permitido poder revisar la importancia de los conceptos extraídos al poder visualizar los párrafos en los cuales aparecen dichos conceptos.

La versión evaluada de Copernic Summarizer corresponde a la 2.0 de Diciembre del 2001.

## 3.3 Nomino

Nomino es un motor de búsqueda desarrollado por la Universidad de Quebec en Montreal, anteriormente llamado Termino, y que actualmente es distribuido por Nomino Technologies [13]. Nomino adopta un enfoque morfosintáctico y es capaz de extraer términos pero también es capaz de identificar conceptos y de aislar sus relaciones semánticas [18]. El analizador morfológico usado por Nomino utiliza una *"lematización"* (*"stemming"* en inglés) lo que significa que el prefijo y el sufijo de una palabra son eliminados para obtener una palabra simple. Esta lematización permite reducir una palabra a su más mínima forma, llamada *"raíz"* [4]. Por ejemplo, si tenemos el concepto siguiente: *"bibliotecas digitales"*, Nomino formará la cadena siguiente: *"biblio digit"*. Así mismo, Nomino aplica ciertos criterios empíricos para filtrar el ruido encontrado en los conceptos extraídos. Estos criterios incluyen la frecuencia y la categoría, además del uso de unas listas que contienen palabras que no deben de ser tomadas en cuenta a la hora de efectuar una extracción. Nomino produce dos tipos de índices interactivos, los cuales contienen todos los conceptos que resumen de manera más precisa el contenido de un documento dado. Uno de los índices creados es muy general ya que contiene todas los conceptos encontrados en el documento. En cambio, el otro índice llamado *"índice prominente"* (en francés llamado: *"index de saillance"*) contiene los conceptos que para Nomino son los más pertinentes para describir un documento dado.

El llamado *"índice prominente"* está basado en dos principios: *"la ganancia para expresar"* y *"la ganancia para alcanzar"*. La *"ganancia para expresar"* clasifica los conceptos de acuerdo a su localización dentro del documento. Por ejemplo, si un párrafo habla sobre un sólo concepto entonces este concepto será clasificado como importante. La *"ganancia para alcanzar"* clasifica los conceptos de acuerdo a la frecuencia. De esta manera, si una palabra es muy rara entonces será

seleccionada como importante. Por ejemplo, en un documento dado, cuando se encuentran los conceptos *"biblioteca digital"* y *"arquitectura de la biblioteca digital"*, sólo el segundo concepto será seleccionado como pertinente ya que es mucho más completo que el primero. Sin embargo, los dos conceptos pueden ser seleccionados como pertinentes al mismo tiempo sólo en el caso de que el primer concepto *"biblioteca digital"* tenga una frecuencia mucho menor que el segundo concepto.

La versión evaluada de Nomino corresponde a la 4.2.22 del 25 de Julio del 2001. Esta versión de evaluación comprende todas las funciones de la herramienta y es posible utilizarla durante 10 horas no consecutivas.

## 3.4 TerminologyExtractor

TerminologyExtractor (TE) es una herramienta distribuida por Chamblon Systems Inc., [17]. El proceso de extracción de términos se lleva a cabo en dos pasos: (1) el paso de *"limpiar el texto"* y (2) el paso para la *"extracción de colocaciones o combinaciones"*.

El objetivo del paso de *"limpiar el texto"* es asegurarse que todas las inconsistencias sean eliminadas del texto de entrada. Para esto, TE utiliza una lista de exclusión que contiene todas las palabras que no son importantes a extraer. Además, TE utiliza un diccionario que es definido por el usuario y que contiene palabras importantes que no están contenidas en el diccionario del sistema.

El paso para la *"extracción de colocaciones o combinaciones"* produce una lista de colocaciones o términos con su respectiva frecuencia de aparición. TE también produce una lista de palabras y *"no-palabras"*. Estas *"no-palabras"* son todas aquellas palabras que no han sido encontradas en ambos diccionarios: el diccionario del sistema y el diccionario definido por el usuario.

La versión evaluada de TerminologyExtractor corresponde a la 3.0.

## 3.5 Xerox Terminology Suite

La herramienta Xerox Terminology Suite (XTS) de la compañía Xerox es un sistema de manejo de terminología que permite la creación de diccionarios multilingües y monolingües a partir de la extracción automática de términos o conceptos [19]. En nuestra evaluación, hemos considerado únicamente dos de los componentes de XTS: *"TermFinder"* y *"TermOrganizer"*.

El componente *"TermFinder"* construye automáticamente una base de datos de términos extraídos. Este componente utiliza herramientas lingüísticas para saber cómo una frase está construida. Así mismo, *"TermFinder"* utiliza una herramienta de extracción de sustantivos, por ejemplo, para saber cuando un sustantivo está seguido de un adjetivo calificativo.

El componente *"TermOrganizer"* se encarga del manejo de la base de datos que ha sido creada por el componente *"TermFinder"*. A través del uso de *"TermOrganizer"* es posible realizar modificaciones y especificaciones de los términos extraídos.

La versión evaluada de XTS corresponde a la 2.0 de Febrero del 2001. Esta versión corresponde a la herramienta llamada *"XTS the Terminology Suite"*.

En la Sección 4, presentamos la metodología utilizada para la evaluación de las herramientas de extracción automática de conceptos.

# 4 Metodología para la evaluación de herramientas de extracción automática de conceptos

La extracción de conceptos o términos está basada en dos diferentes enfoques: *"la asignación de frases claves"* y *"la extracción de frases claves"* [10]. Por el término *"frase clave"* nos referimos a una frase compuesta por dos o más palabras, las cuales describen en general algún tema o concepto tratado dentro de un documento en particular.

El trabajo de *"asignación de frases claves"* se lleva a cabo mediante el uso de un vocabulario controlado. De esta manera, se seleccionan los mejores conceptos o frases que describen un documento.

El trabajo de *"extracción de palabras claves"* se lleva a cabo mediante la selección de los conceptos que se encuentran dentro del documento mismo.

En este artículo presentamos sólo herramientas que llevan a cabo la *"extracción de palabras claves"*, lo cual significa que todos los conceptos que son extraídos siempre aparecen en el cuerpo del documento.

## 4.1 Evaluación de la eficiencia

Para evaluar la eficiencia de las herramientas de extracción de conceptos, decidimos comparar la lista generada de conceptos extraídos por las herramientas contra una lista de conceptos generada manualmente por un experto del área.

El primer paso para la evaluación de la eficiencia de las herramientas está basado en el número de coincidencias entre los conceptos generados por la herramienta y los conceptos generados manualmente por un experto del área. Un concepto es seleccionado como relevante sólo cuando existe una correspondencia entre la misma *"secuencia de radicales"*. Por *"secuencia de radicales"* nos referimos a una secuencia de palabras que aparecen en el mismo orden. Para esto, es importante la utilización de la *"lematización"*, la cual permite la búsqueda de las raíces de las palabras.

El segundo paso para la evaluación de la eficiencia de las herramientas está basado en la frecuencia con la que aparece un concepto dentro de un documento. Uno de los métodos más comunes utilizados para buscar la información pertinente está basado en la selección de aquellos conceptos que tienen una frecuencia alta de aparición. Muy a menudo ocurre que una palabra tiene una frecuencia alta incluso aún no siendo importante para el documento [14]. En este caso, para evitar este problema, en algunas herramientas como TerminologyExtractor la frecuencia alta de las palabras que no son pertinentes es eliminada utilizando una lista de palabras que contiene todas aquellas que no son importantes. Durante la fase de extracción, las palabras son comparadas a esta lista con el objetivo de seleccionar sólo aquellas que son relevantes.

## 4.2 Uso de las métricas de *"precisión"* y *"recuperación"* para clasificar la pertinencia de los términos extraídos

Para evaluar la lista de conceptos generada por cada herramienta, hemos comparado esta lista con una lista de conceptos generada manualmente por un experto del área. Las medidas usadas para la evaluación vienen del área de recuperación de información las cuales son: la *"precisión"* y la *"recuperación"*.

La *"precisión"* es una métrica utilizada para evaluar la capacidad que tiene el método de búsqueda para no obtener documentos no pertinentes. Es el cociente entre el total de documentos pertinentes obtenidos y el de documentos obtenidos. Por lo tanto, puede interpretarse como la probabilidad de que un documento obtenido sea pertinente [16]. En nuestro caso, la *"precisión"* es la proporción de SÓLO los conceptos relevantes recuperados.

La *"recuperación"* es una métrica utilizada para evaluar la capacidad que tiene el método de búsqueda para obtener documentos pertinentes, es decir, considerados útiles por quien hizo la consulta. Es el cociente entre el total de documentos pertinentes obtenidos y el de documentos pertinentes de la base. Por lo tanto, puede interpretarse como la probabilidad de que un documento pertinente sea obtenido [16]. En nuestro caso, la *"recuperación"* es la proporción de TODOS los conceptos recuperados, incluso aquellos recuperados que son irrelevantes.

Para evaluar los conceptos extraídos por las herramientas, es necesario primeramente clasificar los conceptos en la categoría de relevantes o irrelevantes. La evaluación llevada a cabo puede ser resumida en la siguiente tabla, Tabla 1.

| | Concepto clasificado como pertinente por el humano | Concepto clasificado como no pertinente por el humano |
|---|---|---|
| Concepto clasificado como pertinente por la herramienta | *a* | *b* |
| Concepto clasificado como no pertinente por la herramienta | *c* | *d* |

**Tabla 1.** Resumen del método de evaluación.

La variable "*a"* representa el número de conceptos generados por el humano que coinciden con los conceptos extraídos por la herramienta. Las variables "*b"* y "*c"* representan el número de veces o de conceptos en que el humano y la herramienta no concordaron en la fase de clasificación. La variable "*d"* representa el número de veces en que el humano y la herramienta concordaron en evaluar un concepto como irrelevante. De esta manera, las formulas de *"precisión"* y de *"recuperación"* son presentadas a continuación:

$$\text{Precisión} = \frac{a}{a+b}. \quad (1)$$

$$\text{Recuperación} = \frac{a}{a+c}. \quad (2)$$

En la siguiente sección, Sección 5, presentamos los pasos seguidos para efectuar la evaluación de las diferentes herramientas así como los resultados obtenidos

# 5 Resultados de la evaluación de las herramientas de extracción automática de conceptos

En esta sección, describimos la experimentación realizada y los resultados obtenidos en nuestra evaluación. Comenzamos con una descripción de los documentos utilizados como corpus y terminamos con un resumen de los resultados obtenidos.

## 5.1 El corpus

Los experimentos presentados en este artículo están basados en dos diferentes grupos de documentos. El primero corresponde a grandes documentos como lo son las tesis doctorales las cuales contienen más de 150 páginas cada una. El segundo grupo de documentos corresponde a pequeños artículos científicos compuestos de aproximadamente ocho páginas cada uno. Todo el corpus utilizado para la evaluación de las herramientas se encuentra en el idioma francés. Así mismo, tanto las tesis como los artículos científicos corresponden al área de la computación. Lo decidimos de esta manera ya que al analizar otras áreas seriamos incapaces de evaluar la pertinencia de los conceptos extraídos por las herramientas.

En la siguiente tabla, Tabla 2, presentamos el tamaño del corpus utilizado. Cabe recalcar, que en ciertos casos, al utilizar una versión de evaluación de las herramientas nos vimos confrontados al problema del tamaño del corpus. De esta forma, decidimos evaluar cada tesis por separado.

| Nombre del corpus | Número total de documentos | Número total de palabras |
|---|---|---|
| Tesis doctorales | 20 | 1 089 701 |
| Artículos científicos | 5 | 15 864 |

**Tabla 2**. Presentación del corpus utilizado para el análisis de las herramientas.

## 5.2 Diseño de experimentos

El primer paso para la evaluación de las herramientas consiste en la creación de una lista de referencia, la cual nos servirá como base para la comparación de la eficiencia de cada una de las herramientas. Esta lista de referencia contiene las palabras o conceptos claves dados por los autores de cada artículo. Esta primera lista es completada manualmente por un experto del área. La lista de referencia para el grupo de grandes documentos es creada en su totalidad ya que cada tesis no contiene palabras claves dadas por el autor de la misma. De esta forma, cada artículo y cada tesis tiene una lista de referencia con los conceptos que el experto considera que representan cada documento.

El segundo paso para la evaluación de las herramientas consiste en la comparación de los conceptos extraídos por cada herramienta contra los conceptos extraídos manualmente y que se encuentran en la lista de referencia. Un punto interesante en éste paso es la comparación de frases largas ya que éstas sólo son extraídas por algunas de las herramientas evaluadas. Por ejemplo, Nomino es la única herramienta capaz de extraer frases largas (conceptos) compuestas por más de tres palabras. Otro punto importante en éste paso es la visualización de algunos conceptos propuestos como pertinentes por las herramientas, por ejemplo en el caso de Copernic Summarizer que permite visualizar los párrafos donde aparecen dichos conceptos. Además, si la lista de referencias contiene un concepto como por ejemplo: *"ontología"*, es posible que a través de los conceptos extraídos por las herramientas se puedan encontrar conceptos que son mucho más expresivos y representativos, como*: "ontología de representación"* o *"conocimiento de la ontología"*.

El tercer paso es el análisis de los valores resultantes al aplicar las herramientas al corpus (Sección 5.1). Los valores que analizaremos en éste paso son:

- El número total de conceptos extraídos por la herramienta,
- El número total de conceptos extraídos por la herramienta y que aparecen en la lista de referencia creada manualmente,
- El número total de conceptos extraídos por la herramienta y que *no* aparecen en la lista de referencia creada manualmente,
- El número total de conceptos extraídos manualmente y que *no* aparecen en la lista generada por la herramienta.

### 5.2.1 Ejemplo de una evaluación de una tesis digital

Para poder ejemplificar cada uno de los pasos del proceso de evaluación, presentamos a continuación el proceso y los resultados obtenidos en el análisis de una de las tesis[4] del corpus utilizado.

El primer paso para la evaluación consiste en la creación de la lista de referencia para cada uno de los documentos. En nuestro ejemplo, ésta lista de referencia contiene los conceptos que al experto le parecen como pertinentes. Algunos de los conceptos que definimos como pertinentes para la tesis son los siguientes: *"algoritmos genéticos"*, *"aprendizaje automático"*, *"documento científico"*, *"memoria de instancias"*, *"reformulación de preguntas"*, *"sistema de búsqueda"*, *"sistema multiagente"*, *"técnica de aprendizaje"*, etc. Esta lista de referencia está compuesta por 31 conceptos.

Una vez que tenemos ésta lista, el segundo paso consiste en aplicar cada una de las herramientas y comparar en una tabla si los conceptos que se encuentran dentro de la lista de referencia han sido extraídos por cada una de las herramientas. La tabla correspondiente, Tabla 3, quedaría de la siguiente forma una vez aplicada cada una de las herramientas al documento:

| Conceptos de la lista de referencia | XTS | Copernic Summarizer | Terminology Extractor | Nomino |
|---|---|---|---|---|
| Algoritmos genéticos | X | X | X | X |
| Aprendizaje automático | X | | X | |
| Memoria de instancias | X | X | | X |
| Reformulación de preguntas | X | X | X | |
| Sistema de búsqueda | X | | X | X |
| ... | | | | |

**Tabla 3**. Ejemplo de comparación de conceptos extraídos por cada una de las herramientas contra los conceptos que se encuentran en la lista de referencia.

El tercer paso consiste en contar el número de conceptos extraídos al igual que las diferencias entre la lista de referencia. Es así como obtenemos para cada uno de los documentos analizados una tabla como la siguiente (Tabla 4):

[4] Tesis correspondiente al ejemplo dado en la sección 5.2: Jeribi L. "Aide à la recherche documentaire adaptée à l'utilisateur". INSA de Lyon, Francia, 7 de Diciembre, 2001.

| | XTS | Copernic Summarizer | Terminology Extractor | Nomino |
|---|---|---|---|---|
| Número total de conceptos extraídos | 6932 | 99 | 3137 | 71 |
| Número total de conceptos presentes en la lista de referencia (a) | 30 | 9 | 19 | 13 |
| Número total de conceptos ausentes en la lista de referencia (b) | 6906 | 90 | 3118 | 58 |
| Número total de conceptos no extraídos (c) | 1 | 22 | 12 | 18 |

**Tabla 4**. Resultados obtenidos en la evaluación de la tesis.

Por último para poder hacer la comparación entre las herramientas (Sección 5.3) los resultados obtenidos en el tercer paso son aplicados a las métricas de *"precisión"* y *"recuperación"* presentadas en la Sección 4.2. Por ejemplo, para la tesis presentada obtendríamos los siguientes resultados (Tabla 5):

| | XTS | Copernic Summarizer | Terminology Extractor | Nomino |
|---|---|---|---|---|
| **Precisión** | 0.004 | 0.09 | 0.006 | 0.183 |
| **Recuperación** | 0.967 | 0.29 | 0.612 | 0.419 |

**Tabla 5**. Ejemplo de porcentajes de *"precisión"* y *"recuperación"* obtenidos en la evaluación de la tesis.

Para este caso, con los resultados obtenidos podríamos decir que la máxima *"precisión"* es la obtenida por Nomino con un 18%, es decir que un 18% de los conceptos extraídos por Nomino son pertinentes. En cambio, XTS tiene un 96% de probabilidades de que entre los conceptos que extrae estén los pertinentes. Como lo muestra la Tabla 4, XTS es la herramienta que más términos extrae ya que selecciona todos los términos que aparecen dentro del documento analizado.

## 5.3 Comparación de las herramientas

En esta sección presentamos los resultados de la evaluación realizada a las cuatro herramientas seleccionadas. Esta comparación es llevada a cabo utilizando los valores descritos anteriormente en la Sección 4.2, es decir que nos basamos en la *"precisión"* y en la *"recuperación"* para realizar la evaluación.

Una manera para evaluar los resultados, presentados en la Tabla 6, es comparar la eficiencia de cada una de las herramientas contra la eficiencia obtenida al aplicar la herramienta Nomino. Podemos apreciar que la *"precisión"* obtenida por Nomino es mucho más alta que la obtenida por las otras tres herramientas (TerminologyExtractor, XTS, Copernic Summarizer).

| | XTS | Copernic Summarizer | Terminology Extractor | Nomino |
|---|---|---|---|---|
| **Precisión** | 0.028 | 0.339 | 0.068 | 0.834 |
| **Recuperación** | 0.905 | 0.51 | 0.648 | 0.651 |

**Tabla 6.** Resultados en términos de *"precisión"* y *"recuperación"* obtenidos al aplicar las cuatro herramientas al corpus de entrada.

Los resultados de la Tabla 6 muestran que la *"precisión"* más alta es la obtenida por Nomino. En comparación con otras herramientas, Nomino extrae menos conceptos irrelevantes. La *"precisión"* más baja es la obtenida por XTS ya que extrae muchos conceptos que son irrelevantes.

En términos de *"recuperación"* XTS obtiene el mejor valor, sin embargo, uno de los inconvenientes es la gran lista que genera ya que cuando se tienen más de 100 conceptos a evaluar el trabajo manual se vuelve largo y tedioso. Por ejemplo, en la Sección 5.2.1 presentamos el caso de una tesis en la que XTS extrajo 6932 conceptos de los cuales casi todos estuvieron presentes en la lista de referencia. Esto se debe en gran medida a que XTS más bien extrae todas las palabras utilizadas en el documento y que no aparecen en la lista de palabras a no extraer. La herramienta TerminologyExtractor tiene el mismo problema que XTS, ambos extraen muchos términos que en la mayoría de los casos no son conceptos. La herramienta Copernic Summarizer es la herramienta que menos conceptos pertinentes extrae ya que sólo selecciona los conceptos que máximo están formados por dos palabras. Siendo bajo el número total de términos que Copernic Summarizer extrae entonces en la gran mayoría de los casos llega a tener un porcentaje de *"recuperación"* que se acerca a más del 50%.

La herramienta Nomino extrae pocos términos al igual que Copernic Summarizer pero al contrario de ésta última herramienta, Nomino extrae conceptos que muchas veces no están en la lista de referencia pero que son pertinentes a utilizar. De esta forma, consideramos que el uso de Nomino podría servir para que inicialmente con los conceptos que extrae se cree una lista de referencia. Además, Nomino es la única herramienta capaz de extraer conceptos que son más complejos. Por ejemplo, en una de nuestras listas iniciales de referencia el experto escogió el siguiente término como pertinente: *"lenguaje de operacionalización Def*"*. En este caso, sólo Nomino pudo extraerlo.

En la Figura 1, presentamos la diferencia entre la *"precisión"* y la *"recuperación"* obtenida por cada herramienta. En este caso, XTS obtuvo el 2.8% de *"precisión"* contra 90.5% de *"recuperación"*. Esto significa que XTS extrae un alto número de términos y que sólo el 2.8% de todos ellos son conceptos pertinentes. Para la herramienta Nomino obtuvimos un 83% de *"precisión"* contra 65.1% de *"recuperación"* lo cual significa que más de la mitad de los términos extraídos por Nomino son conceptos que aparecen en la lista de referencia y por tanto son pertinentes. TerminologyExtractor obtuvo sólo un 6.8% de *"precisión"* y Copernic Summarizer presenta un 33.9% de *"precisión"*. No obstante que Copernic Summarizer tiene muy buena *"precisión"*, aún queda lejos de la que podemos obtener con Nomino que es de un 83.4%.

En conclusión para nosotros la herramienta que extrae más conceptos pertinentes del total de términos extraídos es Nomino. Esta herramienta también presenta la característica de que a través de los términos extraídos genera una red semántica que podría ser de gran utilidad para la generación de índices.

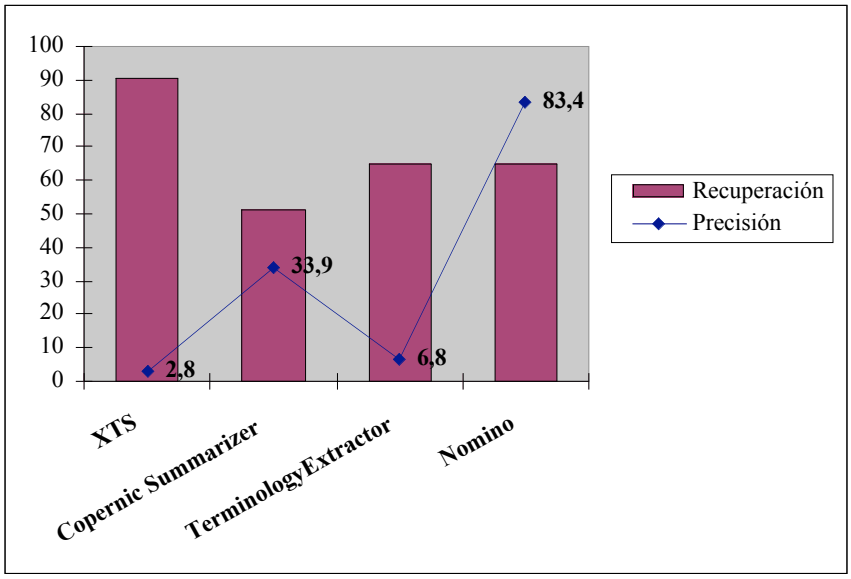


**Fig. 1.** Comparación de los porcentajes obtenidos en términos de precisión y recuperación.

En la Sección 6 presentamos un prototipo de herramienta de anotación para documentos en formato XML.

# 6 Herramienta de anotación

El análisis de cada una de las tesis se basa en el estudio de los capítulos o secciones que contienen una mayor cantidad de conceptos interesantes. Al hacer uso de Nomino, nos hemos dado cuenta que las partes correspondientes a la introducción y a la conclusión no son de gran interés ya que son sólo un resumen de toda la tesis. En este caso cuando analizamos la tesis por completo, Nomino no extrae ciertos conceptos ya que se repiten tantas veces en la introducción y en la conclusión que los toma como si fueran conceptos comunes y no pertinentes. Con el objetivo de extraer el máximo número de conceptos pertinentes hemos eliminado la introducción y la conclusión de las tesis. De igual manera eliminamos la página correspondiente a la portada y la bibliografía. En la Tabla 7 se resumen los resultados obtenidos del análisis de la estructura lógica (generalizando a 5 capítulos). Podemos ver que tanto el índice, la introducción y la conclusión no son de gran importancia, al contrario de los primeros dos capítulos que aportan más del 20% de los conceptos que pueden ser utilizados para caracterizar por completo la tesis .

| Indice | Introducción | Capítulo 1 | Capítulo 2 | Capítulo 3 | Capítulo 4 | Capítulo 5 | Conclusión |
|---|---|---|---|---|---|---|---|
| 8.8% | 7.8% | 20.5% | 22.4% | 17.2% | 18.6% | 18.3% | 9% |

**Tabla 7.** Comparación de la media de porcentajes obtenidos para cada una de las partes que constituyen la estructura lógica de la tesis.

El análisis de la estructura semántica esta basado en la forma en la que el tesista organiza su información. Hay una cierta tendencia a incluir ciertos elementos siempre en la tesis, más cuando se trata de una tesis del área de informática. Por ejemplo, la gran mayoría de las tesis contienen una sección dedicada al *"modelo"* o al *"método"*. El último capítulo es generalmente consagrado al *"prototipo"* o llamado *"estudio de casos"*. En este último capítulo también encontramos una

sección dedicada a la *"arquitectura"* del prototipo. Encontramos que generalmente el capítulo 2 se enfoca al *"estado del arte"*, el cual presenta el tema principal de la tesis así como la investigación que existe a su alrededor. Como podemos ver en la Tabla 7, éste capítulo es el de mayor porcentaje, es decir el que tiene el mayor número de conceptos interesantes y que por si mismos ofrecen bastante información al usuario. Para el análisis semántico, las tesis fueron divididas de acuerdo a los temas tratados. Por ejemplo, *"estado del arte"*, *"métodos"*, *"arquitectura"*, *"modelos"*, *"prototipo"*, *"metodología"*, *"modelización"*, *"experimentación"*, etc. Estos temas pueden aparecer en diferentes secciones o capítulos. Nomino fue utilizado para extraer los conceptos más importantes.

Con el objetivo de hacer uso de los conceptos extraídos de las tesis a partir del uso de las herramientas evaluadas, hemos propuesto una herramienta capaz de *"anotar"* las tesis con los conceptos pertinentes extraídos por Nomino. El objetivo principal consiste en la utilización de un grupo de tesis en formato XML (eXtensible Markup Language) al cual se agregarán automáticamente los conceptos extraídos en forma de etiquetas XML. De esta manera, cuando el párrafo que contiene el concepto es identificado entonces es anotado con una simple etiqueta al principio como *"<nombre-del-concepto>"* y *"</nombre-del-concepto>"* al final. Este esquema de anotación es muy simple y puede ser aplicado fácilmente a un texto usando un editor de XML, sin embargo nuestra herramienta lo hace automáticamente. Los usuarios pueden validar los conceptos propuestos por Nomino y también pueden proponer nuevos conceptos para ser añadidos en forma de etiquetas. Si éstos nuevos conceptos no aparecen dentro del documento entonces son añadidos como etiquetas XML sólo alrededor del resumen de la tesis con el fin de que una vez efectuada una búsqueda, éste párrafo sirva para darle una idea al usuario de la pertinencia de la tesis.

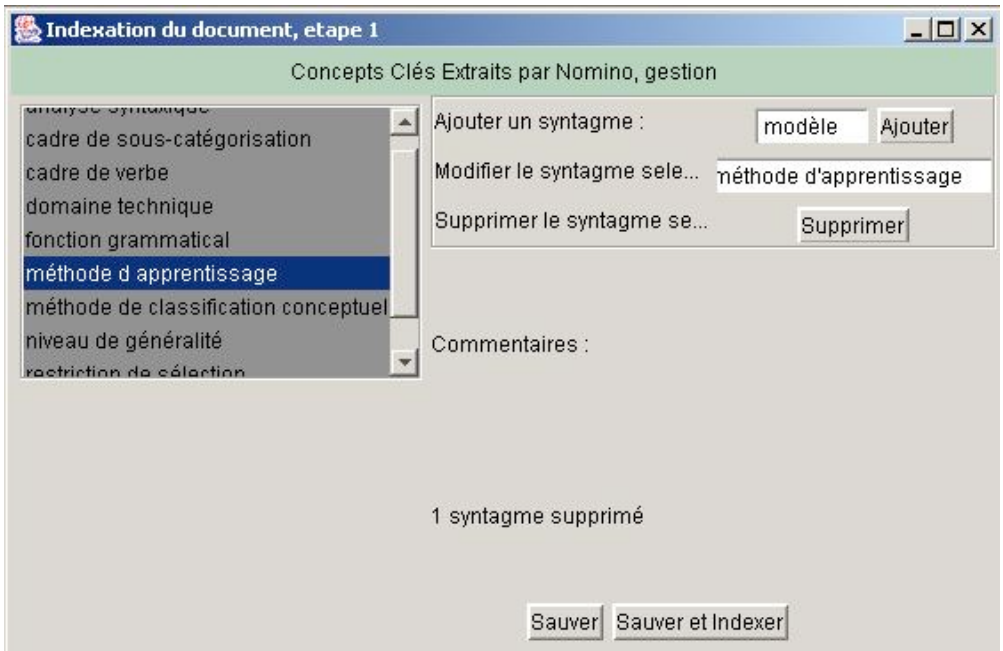


**Fig. 2.** Indexación y validación de los conceptos a través de la herramienta de anotación.

La herramienta de anotación permite el manejo de los conceptos propuestos por Nomino, la indexación y la extracción de los párrafos pertinentes del documento de acuerdo a un cierto criterio de búsqueda. La Figura 2, presenta la herramienta de anotación propuesta. En esta ventana, el

usuario puede visualizar los conceptos propuestos al igual que puede borrarlos o modificarlos. Cuando se realiza una modificación se requiere de una confirmación. Así, las modificaciones son guardadas y los nuevos conceptos son añadidos a la lista.

Los nuevos conceptos generados por el usuario son guardados e incluidos como etiquetas XML dentro del documento inicial. Esto permite la recuperación de los párrafos que contienen la información pertinente a la hora de efectuar una búsqueda (Figura 3).

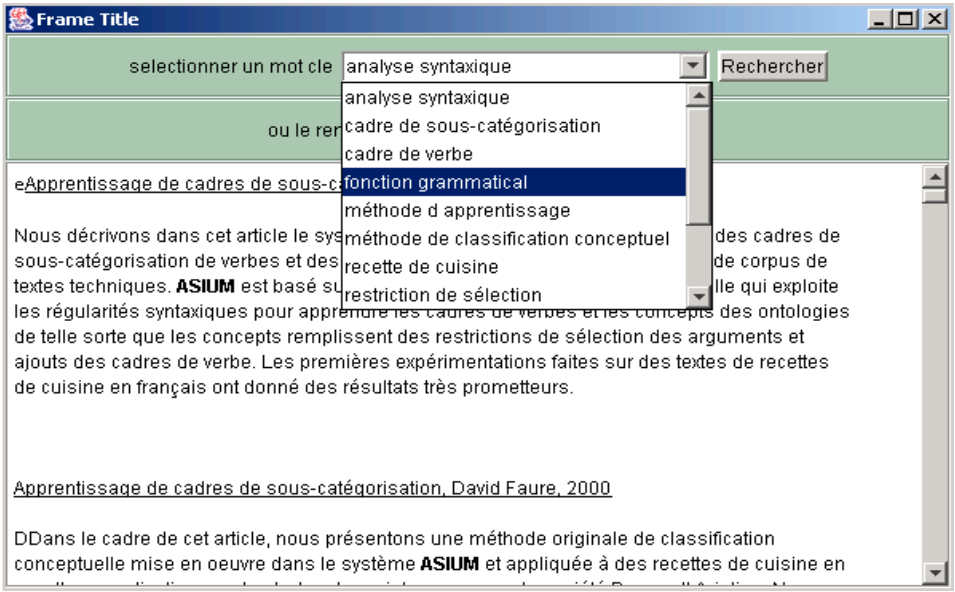


**Fig. 3.** Recuperación de información basada en los conceptos añadidos como etiquetas XML.

Consideramos de gran importancia el uso de XML Schema para estructurar las tesis de acuerdo a un modelo válido. El propósito del estándar XML Schema es definir la estructura de los documentos XML que estén asignados a tal esquema y los tipos de datos válidos para cada elemento y atributo. De esta forma, hemos construido un modelo usando XML Schema que define una tesis de acuerdo a su estructura lógica (introducción, capítulo, párrafo, frase, etc.) y de acuerdo a su estructura semántica (arquitectura, método, modelo, sistema, etc.). No todas las tesis tienen la misma estructura por lo cual, por ejemplo, hemos definido dentro de XML Schema la etiqueta *"<metadato>"* la cual puede aparecer dentro de cualquier frase no importando su ubicación dentro de la tesis. Así mismo, XML Schema nos permite declarar *"<metadato>"* como una etiqueta que puede aparecer cuantas veces sea necesaria. La etiqueta *"<nombre-del-concepto>"* es considerada como una cadena de caracteres correspondiente a la etiqueta *"<metadato>"*.

En la siguiente sección, Sección 7, presentamos las conclusiones y nuestro trabajo futuro.

# 7 Conclusión y trabajo futuro

La búsqueda de la información dentro de una biblioteca digital está focalizada al uso de palabras claves que son construidas a medida que el usuario va obteniendo resultados. Generalmente, el usuario no conoce lo que desea buscar y está esperando que se le propongan opciones.

Dentro de la biblioteca digital CITHER, nuestro trabajo consiste en ofrecerle al usuario la posibilidad de interactuar con la colección de tesis a través de temas que nosotros llamamos *"conceptos"* en lugar de la interacción actual que se lleva a cabo a través del título de la tesis, el nombre del autor, la fecha y el año de edición. Para llevar a cabo ésta interacción, necesitamos herramientas capaces de extraer automáticamente los conceptos ya que manualmente hemos comprobado que es un trabajo largo y tedioso. Con el fin de escoger una herramienta adecuada a nuestras necesidades hemos realizado una evaluación de herramientas de extracción automática de términos.

Este artículo presenta la evaluación de cuatro herramientas de extracción automática de términos o conceptos. Las herramientas evaluadas son: (1) TerminologyExtractor de Chamblon Systems Inc., (2) Xerox Terminology Suite de Xerox, (3) Nomino de Nomino Technologies y (4) Copernic Summarizer de NRC. Algunas de las conclusiones a las que llegamos gracias a los resultados obtenidos por medio de esta evaluación son las siguientes:

- XTS extrae una lista muy grande de términos considerados como pertinentes ya que generalmente extrae todos las palabras compuestas que están bien formadas y que para XTS son términos con el simple hecho por ejemplo de ser un sustantivo seguido de un adjetivo. Sin embargo, éstos términos la gran mayoría de las veces no corresponden a conceptos validos.

- Los conceptos extraídos por Nomino están la mayor parte del tiempo formados por más de dos palabras. Estos conceptos son muy interesantes ya que son mucho más específicos.

- Los conceptos extraídos pueden permitir la evaluación y la agregación de nuevos conceptos que faltaban a la lista inicial de referencia.

- Las herramientas evaluadas pueden ayudar al usuario a describir los temas tratados en un documento a partir del uso de conceptos.

- Finalmente, Nomino es la herramienta más adecuada a nuestras necesidades ya que la gran mayoría de los términos que extrae corresponden a conceptos pertinentes. Al generar una lista corta de términos extraídos, Nomino nos permite una evaluación rápida. Además Nomino extrae *"conceptos complejos"* que las otras tres herramientas son incapaces de extraer.

Los resultados obtenidos en esta evaluación no pueden ser generalizados en otras situaciones de trabajo sin antes hacer un análisis adicional. Dentro de nuestro trabajo, hemos sólo evaluado las capacidades para la extracción automática de conceptos aún cuando algunas de las herramientas evaluadas como lo es Copernic Summarizer realizan otro tipo de tratamientos, sin embargo éstos no estaban contemplados dentro de nuestros objetivos iniciales.

Actualmente estamos trabajando en la concepción de un sistema *"inteligente"* basado en los conceptos extraídos que son modelados como etiquetas XML dentro de cada tesis. Así mismo, estamos creando una ontología basada en los conceptos extraídos por Nomino. Las etapas de la concepción de la ontología y algunos primeros resultados son presentados en [1]. Nuestro sistema *"inteligente"* permitirá ofrecerle al usuario los conceptos apropiados, usando la ontología, para efectuar una búsqueda y así obtener la información pertinente.

Dentro del trabajo futuro consideramos también de gran importancia la introducción de un tesauro capaz de aumentar la pertinencia de las palabras usadas al efectuar una búsqueda gracias al uso de conceptos similares o sinónimos.

## Referencias


[1] R. Abascal, B. Rumpler, J-M. Pinon. Conception d'une Ontologie dans le Contexte d'une Bibliothèque Numérique. ISKO 2003 (International Society for Knowledge Organization), Grenoble, France, July 3-4, 2003.

[2] R. Abascal, B. Rumpler, J-M. Pinon. Improving information retrieval in digital theses using metadata. International Conference on Electronic Publishing (ELPUB 2002). Karlovy Vary, Czech Republic, Elpub 2002 Proceedings pp. 307-316, ISBN 3-897-0035, November 6-8, 2002.

[3] D. Bourigault, C. Fabre. Approche Linguistique pour l'Analyse Syntaxique de Corpus. Cahiers de grammaire 25, pp. 131-151, 2000.

[4] J. Carlberger et al. Improving Precision in Information Retrieval for Swedish using Stemming. 13th Nordic Conference on Computational Linguistics (NoDaLiDa'01), Upsala, May 21-22, 2001.

[5] Copernic Summarizer 2.0, Copernic Technologies Inc, updated in December, 2001. [online] Available at: <http://www.copernic.com/en/products/summarizer/> (24/08/2004).

[6] B. Daille, J. Royauté, X. Polanco. Evaluation d'une Plate-forme d'Indexation de Termes Complexes. Traitement Automatique des Langues (TAL), 41(2), pp. 395-422, 2000.

[7] E. Frank et al. Domain-Specific Keyphrase Extraction, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Morgan Kaufmann, ed., pp. 668-673, ISBN:1-55860-613-0, 1999.

[8] K. Frantzi, S. Ananladou. Automatic Term Recognition using Contextual Cues. Third DELOS Workshop. Cross-Language Information Retrieval. Zurich, Suisse, March 5-7, 1997.

[9] C. Gutwin et al. Improving browsing in digital libraries with keyphrase indexes. Journal of Decision Support Systems, 27, pp. 81-104, 1999.

[10] S. Jones, G. W. Paynter. Human evaluation of Kea, an automatic keyphrasing system. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, June 24-29, 2001, ACM Press, pp.148-156.

[11]S. Jones, S. Lundy, G. W. Paynter. Interactive Document Summarisation Using Automatically Extracted Keyphrases. Proceedings of the 35th Hawaii International Conference on System Sciences, 2002.

[12] M. C. L'Homme. Nouvelles technologies et recherche terminologique, Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. L'impact des nouvelles technologies sur la gestion terminologique, University York, Toronto, August 2001.

[13] Nomino 4.2.22, updated in July 25, 2001. [online] Available at: <http://www.ling.uqam.ca/nomino/> (24/08/2004).

[14] C. Orasan. Building Annotated Resources for Automatic Text Summarization, Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria, España, May 2002.

[15] A. Ribeiro. V. Fresno. A Multi Criteria Function to Concept Extraction in HTML Environment. IC'2001, Las Vegas Nevada, USA. Volume 1, pp. 1-6, 2001.

[16] G. Salton. M. McGill, Introduction to modern information retrieval, McGraw-Hill Book Company, 1983.

[17] TerminologyExtractor 3.0. Chamblon Systems Inc. [online] Available at: <http://www.chamblon.com/terminologyextractor.htm> (24/08/2004).

[18] M. Van Campenhoudt. Les voies de recherche actuelle en terminologie et en terminotique. 7e Université d'Automne en Terminologie, En bons termes, Paris, La Maison du dictionnaire, pp. 109-119, 1998.

[19] Xerox Terminology Suite 2.0. XTS the Terminology Suite, updated in February, 2001. [online] Available at: <http://www.mkms.xerox.com/> (24/08/2004).