Generación de un Método de Expansión de Consultas Basado en Ontologías para un Sistema de Recuperación de Información

M. Rey[1], H. Kuna[1], E. Martini[1], L. Podkowa[1], G. Pautsch[1], E. Zamudio[1]

[1]Dpto. de Informática, Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones.
hdkuna@gmail.com

Resumen. Un Sistema de Recuperación de Información se compone a partir de diferentes módulos, uno de ellos es aquel que permite expandir las consultas ingresadas por el usuario a fin de ampliar el espectro de las búsquedas que se realicen. Existen numerosos métodos para la expansión de consultas, uno de ellos son las ontologías que, por su flexibilidad y capacidad para la representación del conocimiento, constituyen una alternativa viable para determinados contextos. Entre los problemas a los que se enfrentan los procesos de expansión de consultas se encuentra la identificación del lenguaje de la consulta del usuario y la consideración del mismo en las búsquedas y resultados a presentar al usuario. En este trabajo se presenta el desarrollo de un método de expansión de consultas de un meta-buscador para la búsqueda de documentos científicos del área de ciencias de la computación. El método se basa en una ontología de dominio específico e integra con un método para facilitar la realización de búsquedas bilingües.

Palabras clave: ontología, expansión de consultas, meta-buscador, tratamiento del lenguaje.

1 Introducción

En esta sección se describen los ejes teóricos del presente trabajo: Sistemas de Recuperación de Información, Ontologías, Expansión de Consultas y el Tratamiento del lenguaje necesario en este tipo de sistemas.

1.1 Sistemas de Recuperación de Información

Un Sistema de Recuperación de Información (SRI) se define como un proceso con capacidad para recuperar, almacenar y retornar información ante las necesidades de un usuario [1, 2]. Existen implementaciones de SRI que trabajan en la web sobre contextos generales o particulares, integrando diversas técnicas con el objetivo de mejorar la calidad de los resultados que presentan al usuario final [3]. Un meta-buscador es una variante de SRI que se construye modularmente definiendo cada uno de sus componentes en forma específica para su funcionamiento en un contexto

particular [3]. En trabajos anteriores los autores han planteado la estructura general y los componentes complementarios de un meta-buscador que opera sobre documentos científicos del área de ciencias de la computación [4]. El mencionado SRI es al cual se pretende integrar el método de expansión de consultas que se propone en el presente trabajo, que se puede observar destacado en la figura 1.

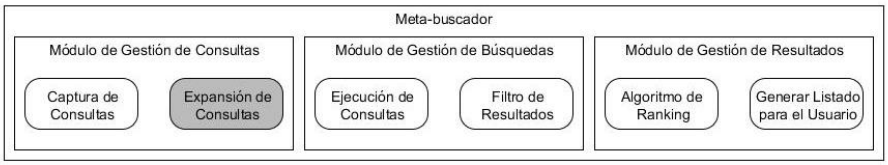

Fig. 1. Estructura del SRI, destacando el componente que se genera.

1.2 Ontologías

En un sentido general, una ontología puede ser definida como una forma de representación del conocimiento de un ámbito específico, se construye utilizando tanto su terminología propia como las relaciones específicas que puedan definirse entre sus conceptos [5, 6]. Adaptando la definición al área de ciencias de la computación se puede definir a una ontología como un esquema conceptual de un dominio de conocimiento, que brinda la posibilidad de transmitir información entre sistemas tanto interna como externamente [7]. De esta manera se presenta como una herramienta de gran valor para actividades de recuperación de información, permitiendo el tratamiento y el análisis del conocimiento que representa en forma automática, a través del sentido que brindan las relaciones, propiedades y reglas que se definen entre las clases e instancias que simbolizan los conceptos del área de conocimiento a representar [8].

1.3 Expansión de consultas en un SRI

Entre las operaciones de un SRI que se consideran adaptables a su contexto de aplicación se destaca el tratamiento que el mismo realiza sobre las consultas que ingresa el usuario [9, 10]. Un tipo de tratamiento es la expansión de consultas, que permite la incorporación de diversos términos a la consulta original con el objetivo de ampliar el espectro de búsqueda y de documentos a los que pueda acceder el SRI. Como resultado se obtiene un nuevo conjunto de consultas con términos adicionales, denominadas expansiones [11, 12], estas nuevas consultas serán ejecutadas sobre las fuentes de datos del SRI generando conjuntos de resultados que luego serán unificados y procesados, generando un único listado para el usuario. En la literatura relacionada con la recuperación de información en la web, se identifican diferentes métodos para realizar expansión de consultas, cada uno de ellos haciendo uso de técnicas y herramientas diferentes como ser: tesauros, diccionarios, sistemas expertos, entre otros [11–14].

Se ha propuesto un método de expansión de consultas basado en la utilización de una ontología de dominio específico para el meta-buscador antes mencionado, en el que se utilizan conceptos propios de un área temática que sea seleccionada por el usuario [15]. En el presente trabajo se pretende mejorar el método a partir de la incorporación de nuevos componentes que incluyan el tratamiento del lenguaje en el que el usuario realiza su consulta.

1.4 Tratamiento del lenguaje en un SRI

Uno de los problemas en la construcción de un método de expansión de consultas para un SRI, es el tratamiento que se realice del lenguaje en el que el usuario escribe las consultas [16]. Entre las posibles soluciones a este problema se destacan aquellas que operan en forma automática para realizar una traducción [17], específicamente se reconocen cuatro estrategias para el control de las traducciones:

- Emparejamiento de términos entre consultas y documentos sin traducción.
- Traducción de las consultas al idioma de los documentos.
- Traducción de los documentos al idioma de las consultas.
- Traducción de los documentos y las consultas a un lenguaje común.

En el contexto del presente trabajo, al tratarse de un meta-buscador que no realiza un tratamiento interno de los documentos a presentar al usuario, la traducción de las consultas al idioma de los documentos se reconoce como el enfoque más adecuado [18], ya que permitiría solucionar algunos problemas en la recuperación de documentos, como ser: la existencia de bases de datos que no procesan consultas en idioma castellano o contienen únicamente documentos en inglés y otras en las que la consulta enviada no es traducida y por lo tanto, aunque contienen documentos en ambos idiomas, se acota el espectro de la búsqueda retornando escasos resultados.

Por lo mencionado previamente se plantea un método de expansión de consultas que en un primer paso reconozca el idioma del texto introducido por el usuario, y en base a la detección, modifique el proceso de expansión haciendo un uso diferente de la ontología. Esto con el objetivo de incrementar la cobertura de las búsquedas a ejecutar y consecuentemente mejore la cantidad de los resultados a obtener, incluyendo aquellos que no se encuentren en el idioma original de la consulta.

2 Construcción de la ontología

La construcción de una ontología se puede basar en diferentes metodologías [19, 20], en el presente trabajo se ha optado por la que mejor se ajusta al objetivo de su desarrollo, facilitando su implementación, sus pasos se pueden observar en la figura 2.

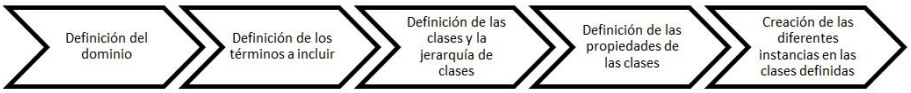

Fig. 2. Pasos para la construcción de la ontología.

Definición del dominio de la ontología.

En el caso del presente trabajo, el objetivo de la ontología consistió en formar una base sobre la cual se pudiera desarrollar un método de expansión de consultas para un meta-buscador de documentos científicos del área de ciencias de la computación. Por lo tanto el dominio de la ontología se acotó a cada una de las subáreas temáticas en las que puede dividirse la disciplina [21].

Se planteó comenzar por la subárea de Inteligencia Artificial (IA), de manera tal de contar con una primera ontología que pudiera ser evaluada individualmente y como base del método de expansión de consultas del SRI, para posteriormente desarrollar las ontologías del resto de las subáreas temáticas de las ciencias de la computación.

Definición de los términos a incluir en la ontología.

En esta instancia se realizó un análisis sobre el estado del arte de la disciplina [22–24], de modo de generar un listado de términos conformado por los conceptos que definen y caracterizan a las diferentes subáreas en las que puede dividirse la IA, palabras clave de cada una de ellas, glosarios específicos, sinónimos de cada término encontrado, entre otros. Con el objetivo que el meta-buscador pudiera expandir consultas tanto en castellano como en inglés, los términos seleccionados fueron traducidos, y se determinó construir una ontología en cada idioma, agregando en cada una la traducción de sus conceptos al idioma inverso.

Definición de las clases y la jerarquía de clases.

Para la definición de la jerarquía de clases se utilizó el método top-down, dado que se consideró como la opción más apropiada para representar la taxonomía innata a los conceptos utilizados. En primera instancia se definieron cuáles conceptos del listado generado en el paso anterior pasarían a convertirse en clases y cuáles en instancias de tales clases. En este sentido, aquellos conceptos que representaban categorías en las que se subdivide la disciplina se transformaron en clases, incluyendo la mayoría de ellos, instancias que se consideraron como subclases.

Definición de las propiedades de las clases.

En un siguiente paso se procedió con la definición de las propiedades de cada clase de la ontología. Una propiedad se define como un atributo que permite describir a un concepto, incluyendo las relaciones que el mismo pueda mantener con otros elementos de la ontología. En el contexto del presente trabajo se definieron dos tipos de relaciones, implícitas y explícitas. Aquellas que están dadas por la jerarquía de clases y en las que no interviene ningún elemento extra se denominan implícitas. En este caso se definieron las siguientes relaciones:

- “padre – hijo”: representando que una clase “contiene” a otra clase o concepto.
- “hijo – padre”: representando que una clase o concepto “es contenida por” otra clase.
- “hermano”: representando que *n* clases o conceptos comparten un mismo “padre”.

Por otra parte, las relaciones que son definidas considerando el contexto de utilización de la ontología se denominan explícitas, en este caso particular se definieron dos relaciones de este tipo:

- "Sinónimo": se trata de una relación de igualdad de significado entre conceptos.
- "Traducción": se trata de una relación-propiedad en la que a cada clase o concepto se lo relaciona con su traducción al idioma inverso al principal de la ontología.

Se han dejado de lado otros componentes de una ontología como son los axiomas y las reglas, ya que su funcionalidad no se consideró necesaria en el contexto en el que se va a usar la ontología.

Creación de las diferentes instancias en las clases definidas.

Para finalizar la construcción de la ontología se definieron cuáles conceptos pasarían a ser las instancias de cada una de las clases definidas en la jerarquía. Los términos utilizados fueron aquellos de mayor atomicidad, obtenidos desde el listado de conceptos conformado en el segundo paso del proceso de construcción.

Implementación en una herramienta software.

Para llegar a implementar un método de expansión de consultas basado en las ontologías diseñadas se utilizó una herramienta software que permitiera su exportación a diferentes formatos, para que posteriormente pudiera ser integrada al SRI. En este caso se optó por el software Protègè [25], que permitió exportar la ontología a un formato OWL[1] y para poder hacer uso de la ontología desde el SRI se utilizó el framework Jena [26] que permitió la realización de consultas sobre la ontología para utilizar su contenido en la expansión de consultas del meta-buscador.

3 Método para la identificación del lenguaje de las consultas

El método que se planteó se basó en la medida de la similitud que pueda existir entre el texto de la consulta ingresada por el usuario y un texto de referencia para la definición de cada idioma almacenado internamente en el meta-buscador, en este caso castellano e inglés, estrategia que se reconoce en otras publicaciones de la temática [27, 28]. Para el procesamiento del texto se debió seleccionar un modelo de representación común para el texto de la consulta y del texto de referencia, optando en este caso por un modelo vectorial.

3.1 Modelo de representación

A fin de poder realizar las comparaciones entre los textos de la consulta y el de referencia para cada idioma se debió definir un modelo de representación común para poder realizar una comparación efectiva entre ambos. El modelo de representación vectorial es utilizado en varias publicaciones del ámbito de recuperación de

[1] OWL es acrónimo (en inglés) para Web Ontology Language.

información [3, 29]. En este modelo cada documento es almacenado en un vector en un espacio n-dimensional, donde cada una de las posiciones del vector guarda valores relacionados con las ocurrencias de diferentes componentes del documento, pudiendo ser: términos, unigramas, n-gramas[2], entre otros; de esta manera, la posición *i* de un vector almacena el valor de la frecuencia de ocurrencia del componente *i* que se haya definido. Al explotar este tipo de representación se pueden efectuar comparaciones entre documentos [27, 28], y determinar el idioma de un texto objetivo, comparando su vector contra uno de referencia de cada idioma a detectar y midiendo su similitud.

Para el contexto del presente trabajo se ha tomado que cada componente del vector contendrá unigramas y n-gramas, con n = 2. En base a la frecuencia de aparición de estos elementos se planteó la comparación entre documentos, midiendo únicamente la frecuencia de los componentes presentes en el texto.

3.2 Métrica para medir la similitud entre documentos

Teniendo los vectores de los textos de referencia y el correspondiente a la consulta del usuario, se debe contar con alguna métrica que permita determinar la similitud entre los valores de los vectores antes mencionados. En el presente trabajo se optó por utilizar la ecuación del coseno del ángulo entre los vectores implicados, cuyo cálculo se define a partir de la ecuación 1.

$$\cos \theta = d_1 * d_2 / |d_1| |d_2| \quad (1)$$

En la ecuación 1, d1 y d2 son los vectores correspondientes al texto de la consulta del usuario y del texto de referencia de cada idioma respectivamente, el * simboliza el producto escalar de los vectores y |d| es el módulo del vector d.

De esta manera se cuenta con una métrica para determinar la similitud entre los documentos, pudiéndose generar con estos componentes el método para la detección del idioma requerido por el SRI.

3.3 Construcción del método

Haciendo uso de los componentes antes descriptos se generó el método para la detección del idioma de la consulta que ingresa el usuario al meta-buscador. El proceso, inicia a partir de la captura del texto de la consulta y su representación a través del modelo vectorial. Continúa con la comparación contra los vectores correspondientes a los idiomas de referencia del SRI, castellano e inglés, usando la ecuación 1, comparando al vector de la consulta con los vectores de los dos idiomas, que se mantienen almacenados internamente en el meta-buscador.

A partir de una comparación entre los valores obtenidos, aquel que resulte más alto será el que determine el idioma en el que el usuario haya realizado su consulta, pudiendo continuar desde esa instancia con el proceso de expansión de consultas del SRI.

---

[2] Un n-grama es una secuencia de n letras del alfabeto definido.

4 Desarrollo del método de expansión de consultas

El método para la expansión de consultas se dividió en dos etapas, inicialmente se busca detectar el idioma de la consulta del usuario para posteriormente utilizar la ontología correspondiente según el idioma para buscar aquellos conceptos que guarden mayor relación con la consulta ingresada para realizar la expansión.

Para la primera etapa se utiliza el método descripto en la sección 3, capturando la consulta del usuario, en adelante *"consulta_original"*, y determinando su idioma, a continuación se instancia la ontología correspondiente al idioma determinado y se pasa a la siguiente etapa del proceso de expansión.

En la segunda etapa se busca dentro de la ontología aquel o aquellos conceptos que sean más similares a la *consulta_original*, y se utilizan las relaciones y propiedades de los mismos para generar las expansiones. La búsqueda en la ontología del concepto más similar a la *consulta_original* se compone de los siguientes pasos:

1. Por cada término de la *consulta_original*: se recorren los elementos de la ontología, clases e instancias, buscando uno o más conceptos con la mayor cantidad de coincidencias sintácticas con el término, almacenando cada uno de los resultados en una colección auxiliar.
2. Se examina la colección obtenida como resultado del paso anterior:
    a. En caso de que se encuentre vacía, se finaliza la expansión sin resultados.
    b. Si la colección contiene un único elemento, el mismo pasa a ser el "*término_candidato*" a partir del cual se realizará la expansión.
    c. En caso de que la colección contenga *n* elementos se analiza en forma individual a cada uno, seleccionando al candidato en base a las coincidencias sintácticas con la *consulta_original*. Si en este análisis se obtienen cantidades iguales para dos o más elementos se decide a partir de las relaciones en la ontología de tales elementos:
        i. Si todos los conceptos están contenidos en una misma clase, se selecciona al concepto "padre" como *término_candidato*.
        ii. En caso de que los conceptos no estén contenidos en una misma clase, se tomará como *término_candidato* al "padre" que referencie una mayor cantidad de instancias.
        iii. Si además los padres poseen la misma cantidad de instancias referenciadas se genera una colección de *términos_candidatos*.

Como resultado de la búsqueda se cuenta con uno o más candidatos para llevar a cabo el proceso de expansión, el mismo se compone de los siguientes pasos[3]:

1. A través de las relaciones definidas en la ontología se obtiene el concepto "padre" del candidato y se lo pasa a denominar "*concepto_padre*".
2. Se obtienen los conceptos que sean del mismo nivel, es decir los "hermanos" del *término_candidato*, que se almacenan en la colección "*conceptos_hermanos*[]".

[3] En caso de trabajar con un conjunto de *términos_candidatos* se aplican los mismos pasos para cada uno de ellos.

3. Se obtiene la colección de sinónimos del *término_candidato*, en caso de existir, generando la colección "*sinónimos_concepto*[]".
4. Se obtiene la traducción de cada uno de los elementos obtenidos en los pasos anteriores, almacenándolos en una colección denominada "*traducciones*[]".
5. Se generan las expansiones de la consulta del usuario utilizando los elementos obtenidos en los pasos anteriores, inicialmente en el idioma original de la consulta:

- Expansión_1 = *consulta_original* AND *término_candidato*
- Expansión_2 = *término_candidato* AND *concepto_padre*
- Expansión_3 = *término_candidato* OR *conceptos_hermanos*[]
- Expansión_4 = *término_candidato* OR *sinónimos_concepto*[]

De la misma manera se generan las expansiones en el idioma opuesto al de la consulta:

- Expansión_traducida_1 = *traducciones*[*candidato*] AND *traducciones*[*padre*]
- Expansión_traducida_2 = *traducciones*[*candidato*] OR *traducciones*[*hermanos*]
- Expansión_traducida_3 = *traducciones*[*candidato*] AND *traducciones*[*sinónimos*]

Como resultado del proceso completo de expansión de consultas se cuenta con dos conjuntos de expansiones de consultas, que serán utilizados para ejecutar búsquedas sobre las fuentes de documentos definidas en el SRI, considerando el idioma principal en el que sea preferente la realización de la consulta. De esta manera se cuenta con un conjunto de consultas en ambos idiomas a modo de solución del problema planteado para el presente trabajo, constituyendo una herramienta para expandir consultas a partir de conocimiento de la disciplina de los documentos científicos a recuperar.

4.1 Validación del método de expansión de consultas

La validación del método propuesto se planteó para su realización con la colaboración de un grupo de expertos en el área de IA, ya que se debió determinar si la expansión sería de utilidad para expandir el espectro de las búsquedas a ejecutar en el SRI. Para llevar a cabo la experimentación se han realizado diversas consultas específicas documentando los resultados del proceso de expansión para su evaluación por parte de cada experto, quienes determinaron un valor entre 1 y 10 como medida de calidad de la expansión. Los resultados obtenidos pueden observarse en la tabla 1.

En general los expertos han evaluado los resultados de la expansión como positivos, con la salvedad de algunos casos en los que han considerado que los términos incluidos por las relaciones representadas en la ontología podrían generar una amplitud excesiva de la consulta cuando se trate de consultas cuyos términos candidatos sean clases del primer nivel de la ontología, es decir, aquellas cuyo concepto padre sea directamente la raíz de la ontología. Siendo esta situación un problema a solucionar para futuras versiones del método.

Tabla 1. Resultados de la validación realizada por los expertos

Consulta realizada	Efectividad promedio
agentes inteligentes AND recuperación de información	6.2
search methods AND deep first search	7.4
unsupervised learning AND backpropagation networks	6.8
algoritmos genéticos OR algoritmos evolutivos	7
fuzzy sets AND expert systems	8.2

5 Conclusiones y trabajos futuros

Se ha generado un método para la expansión de consultas a realizar por un meta-buscador que opera recuperando documentos científicos del área de ciencias de la computación. La expansión se realiza agregando contenido para contextualizar la consulta a través de una ontología de dominio específico, generada para un subárea de las ciencias de la computación. Además se generó un método para la detección del idioma en el que se realiza la consulta y a partir de tal detección se efectúan ajustes en el método de expansión para poder ejecutar búsquedas en inglés y castellano. De esta manera se ha planteado una solución a los problemas abordados en el presente trabajo y la misma ha sido validada positivamente por un conjunto de expertos del área.

Como trabajos a futuro se pueden mencionar: generar las ontologías restantes para el resto de las subáreas temáticas de la disciplina para completar la expansión de consultas; determinar mejoras a incorporar en el proceso de expansión y determinar si es necesario realizar cambios en el método de detección del idioma.

6 Bibliografía

1. Salton, G., Mcgill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc. (1983).
2. Kowalski, G.: Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers, Norwell, MA, USA (1997).
3. Olivas, J.A.: Búsqueda Eficaz de Información en la Web. Editorial de la Universidad Nacional de La Plata (EDUNLP), La Plata, Buenos Aires, Argentina (2011).
4. Kuna, H., Rey, M., Martini, E., Solonezen, L., Podkowa, L.: Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. Rev. Latinoam. Ing. Softw. 2, 107–114 (2013).
5. Grubbs, F.E.: Procedures for Detecting Outlying Observations in Samples. (1974).
6. Hendler, J.: Agents and the semantic web. IEEE Intell. Syst. 16, 30–37 (2001).
7. Muñoz, A., Aguilar, J.: Ontología para bases de datos orientadas a objetos y multimedia. Av. En Sist. E Informática. 6, 167–184 (2009).
8. Sánchez López, S.E.S.: Modelo de indexacion de formas en sistemas VIR basado en ontologias, (2007).

9. Ruiz-Morilla, J., Serrano-Guerrero, J., Olivas, J., Viñas, E.: Representación Múltiple de Consultas: Una alternativa a la Expansión de Consultas en Sistemas de Recuperación de Información. Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF. pp. 531–536 (2010).
10. Alsaffar, A.H., Deogun, J.S., Raghavan, V.V., Sever, H.: Enhancing Concept-Based Retrieval Based on Minimal Term Sets. J. Intell. Inf. Syst. 14, 155–173 (2000).
11. Villa, M. de la, García, S., Maña, M.J.: ¿De verdad sabes lo que quieres buscar? Expansión guiada visualmente de la cadena de búsqueda usando ontologías y grafos de conceptos. Proces. Leng. Nat. 47, 21–29 (2011).
12. Chang, Y., Ounis, I., Kim, M.: Query reformulation using automatically generated query concepts from a document space. Inf. Process. Manag. 42, 453–468 (2006).
13. Gauch, S., Smith, J.B.: An expert system for automatic query reformulation. J. Am. Soc. Inf. Sci. 44, 124–136 (1993).
14. French, J.C., Brown, D.E., Kim, N.-H.: A classification approach to Boolean query reformulation. J. Am. Soc. Inf. Sci. 48, 694–706 (1997).
15. Kuna, H., Rey, M., Podkowa, L., Martini, E., Solonezen, L.: Expansión de Consultas Basada en Ontologías para un Sistema de Recuperación de Información. Presented at the XVI Workshop de Investigadores en Ciencias de la Computación (2014).
16. Ballesteros, L., Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 84–91. ACM, New York, NY, USA (1997).
17. Oard, D.W.: A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In: Farwell, D., Gerber, L., and Hovy, E. (eds.) Machine Translation and the Information Soup. pp. 472–483. Springer Berlin Heidelberg (1998).
18. Oard, D.W.: Alternative approaches for cross-language text retrieval. AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence (1997).
19. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series. Facultad de Informática (UPM), Stanford University, EEUU (1997).
20. Noy, N.F., Mcguinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. (2001).
21. ACM: The ACM Computing Classification System (CCS), (2012).
22. Feigenbaum, E.A., Barr, A., Cohen, P.R.: The handbook of artificial intelligence. Addison-Wesley New York (1989).
23. Rich, E., Knight, K.: Artificial intelligence. McGraw-Hill New. (1991).
24. Nilsson, N.J.: Principles of artificial intelligence. Springer (1982).
25. Stanford Center for Biomedical Informatics Research: Protègè. Stanford University (2014).
26. Apache Software Foundation: Jena. (2014).
27. Bastrup, S., Pöpper, C.: Language detection based on unigram analysis and decision trees. Proj. 2003. 27 (2003).
28. Russell, G., Lapalme, G., Plamondon, P.: Automatic Identification of Language and Encoding. Rapp. Sci. Lab. Rech. Appliquée En Linguist. -Form. RALI Univ. Montr. Can. 7–2003 (2003).
29. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge university press Cambridge (2008).