

FOG: ARQUITECTURA FLEXIBLE PARA LA GENERACIÓN AUTOMÁTICA DE ONTOLOGÍAS

Emilio F. Viñas¹, Jesús Serrano-Guerrero², José A. Olivas², Javier de la Mata², Andrés Soto³,

Departamento de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha,
Paseo de la Universidad 4, 13071, Ciudad Real

¹ fdez.emilio@gmail.com

² jesus.serrano, joseangel.olivas, javier.mata@{uclm.es}

³ soto_andres@yahoo.com

Resumen

Tras realizar una consulta a un sistema de recuperación de información (SRI), se obtienen datos de los cuales, sólo algunos tienen relación con la consulta y otros datos relacionados contextualmente con la consulta no son recuperados. Una de las causas de este problema es que los sistemas de recuperación de información actuales no conocen qué relaciones guardan unas palabras con otras y, por tanto, no son capaces de analizar el contexto del dominio en el que se formuló la consulta.

Una posible solución al problema planteado anteriormente es la creación de ontologías que establezcan relaciones entre palabras creando conceptos capaces de definir el dominio. En este artículo se presenta la arquitectura de una plataforma llamada Fuzzy Ontology Generator (FOG). Esta plataforma establece las relaciones existentes entre palabras generando conceptos que definan un dominio de forma automática.

Palabras Clave: Ontología, Lógica borrosa, Sistema de Recuperación de Información.

1 INTRODUCCIÓN

Un sistema de recuperación de información, se basa en la representación, almacenamiento, organización y acceso a la información. Un SRI almacena información mediante la indexación de datos y cuando el usuario del SRI expresa una necesidad, el sistema establece un ranking de elementos en base a dicha consulta.

Formalmente, un SRI [1] se puede describir como una cuádrupla $SRI = [D, Q, F, S]$ en la cual:

- D : es el conjunto de documentos que representa a los elementos almacenados.

- Q : es el conjunto de consultas que representan la consulta del usuario al sistema.
- F : es la herramienta que modela la representación de documentos, las consultas del usuario y sus relaciones.
- $S: Q \times D \rightarrow \mathcal{R}$ es una función que expresa como de relevante es un documento respecto a la consulta de un usuario.

Más importante que la cantidad de datos que se puedan almacenar en un sistema de recuperación de información, es la obtención de datos útiles tras una consulta a dicho sistema. Existen sistemas de recuperación de información basados en conocimiento, que aumentan la eficacia de la recuperación de información y dan más flexibilidad a la consulta.

La incorporación de conocimiento a los sistemas de recuperación de información provoca un salto cualitativo y cuantitativo en la búsqueda de información. La extracción de conocimiento y su posterior incorporación a un sistema de recuperación de información, se basa en el trabajo de grupos de expertos que extraen conceptos y establecen relaciones entre ellos. En el caso de que el dominio a modelar sea amplio, se requiere de un gran esfuerzo que convierte la obtención de las reglas que forman el conocimiento, en una tarea imposible de abordar. Para poder añadir a los sistemas de recuperación de información la capacidad de analizar el contexto de una consulta, bastaría con la incorporación de un conocimiento que describiese las relaciones existentes entre los términos que forman el dominio.

En este artículo se presenta FOG, una herramienta capaz de crear ontologías de forma automática. La idea es desarrollar una herramienta que permita la posibilidad de albergar distintas estrategias para la generación de ontologías aplicando distintos algoritmos a la hora de establecer las relaciones y atribuirle un grado de similitud semántica a dichas relaciones mediante técnicas de lógica borrosa. Todas estas variantes en el momento de decidir la estrategia a seguir proporcionan a la herramienta una gran flexibilidad.

1. 1. ESTADO DEL ARTE

Son muchos los trabajos propuestos con el objetivo de poder modelar la información y que ésta pueda ser utilizada por máquinas. Este es el caso de Lucarella et al. [6], que en 1991, presentó un ejemplo de un sistema borroso de recuperación de información basado en conocimiento, en el que se aplicaba la teoría de conjuntos borrosos de Zadeh [10]. En este sistema, el conocimiento se estructuraba mediante una red de conceptos. Pero dicha red de conceptos fue establecida por expertos, haciendo la creación de este sistema inviable en el caso de dominios amplios. Esta problemática contribuyó a la idea de extraer el conocimiento de una forma automática.

En 1999, Liang [5] presentó un trabajo en el cual se creaba una red de conceptos de tres niveles de forma automática. En esta red se representan los documentos en el nivel inferior de la red conceptual y el resto de los niveles representan a los conceptos.

Tanto Lucarella como Liang, entre otros autores, formulaban que las relaciones entre nodos fuesen únicas y fijas. En 2003, Horng et al. [4], propuso que las relaciones entre conceptos dependen de la perspectiva, es decir, según el contexto en el que se formule una pregunta, se puede elegir una relación u otra entre un par de nodos. Horng, establece cuatro tipos de relaciones entre nodos: relación borrosa positiva, relación borrosa negativa, relación borrosa generalización y relación borrosa de especificación. A cada una de estas relaciones se les atribuye un valor, proporcionando así semántica a la unión entre nodos.

Como resultado de la extracción automática de conocimiento, se propusieron herramientas como FOGA [9], que genera ontologías de forma automática haciendo uso de lógica borrosa. Gracias a las ontologías se puede representar la semántica entre las palabras de una forma entendible por las máquinas. Según Gruber [3], una ontología es una especificación explícita de una conceptualización. Para nuestra arquitectura, se puede entender una ontología como una cuádrupla $O = [C, P, R, A]$ donde:

- C: es el conjunto de conceptos definidos en el dominio. Un concepto es a menudo considerado como una clase en una ontología.
- P: es el conjunto de propiedades de un concepto. Una $p \in P$ es definida como un instancia de una relación ternaria de la forma $p(c, v, f)$, donde $c \in C$ es un concepto de una ontología, v es una propiedad asociada con c y f define restricciones sobre v .
- R: es un conjunto de relaciones semánticas entre dos conceptos en C.

- A: es el conjunto de axiomas que definen la ontología.

Existen multitud de técnicas para el modelado de información. Por ejemplo, Gomide et al. [7], expone dos teorías. Una basada en la coocurrencia de términos, llamada *Keyword Connection* y otra basada en las relaciones múltiples, *Multirelationship Fuzzy Concepts*. Estas dos teorías, entre otras, ponen de manifiesto la multitud de estrategias para el modelado de información usando lógica borrosa.

1. 2. MOTIVACIÓN

La finalidad de FOG es la de crear ontologías que puedan ser utilizadas por un sistema de recuperación de información. La idea de este proyecto surge como elemento complementario para el framework BUDI [8], desarrollado en colaboración entre la empresa CSD¹ (Computer for Software and Development), con el fin de cubrir algunas de sus deficiencias como es la gestión de información externa útil para los distintos procesos que requieren de cierta información con carga semántica.

En los siguientes apartados de este documento se detalla más detenidamente el funcionamiento completo de la arquitectura de FOG. Más concretamente, en el apartado 2 se hará un repaso de la arquitectura completa de la herramienta y se comentará la parte de configuración de la que se hará uso en el flujo de ejecución de la aplicación, a continuación, en el apartado 3, se analizará cómo se realiza la extracción de datos, cuyo resultado pasará a la fase de modelado de la información de la que se hablará en el apartado 4 y, por último, en lo referente al funcionamiento de FOG, en el apartado 5, se explicarán las distintas representaciones que pueden hacerse sobre el grafo conceptual generado. Finalmente, algunas conclusiones y trabajos futuros serán comentados brevemente.

2 ARQUITECTURA

Se puede definir FOG (Fuzzy Ontology Generator) como una herramienta abierta a la incorporación de nuevos algoritmos con los que se puedan diseñar distintas estrategias para la creación de ontologías de forma automática.

El objetivo es poder realizar variaciones en las funcionalidades e incluso intercambiarlas para poder obtener distintos resultados representados visualmente mediante un grafo o en un lenguaje para la representación

¹ <http://www.csd-systems.com/>

de ontologías (RDF², OWL³, Fuzzy OWL [2]). Así queda patente que es una herramienta en continuo desarrollo al ser posible realizar mejoras de rendimiento, aumentar la funcionalidad de componentes e incorporar nuevas estrategias. FOG debe ser totalmente independiente de cualquier librería externa que pueda ser susceptible de ser utilizada (indexadores, procesadores léxicos o sintácticos, visualizadores...), de tal forma que cualquier componente pueda ser intercambiado sin problemas.

Antes de comentar los principales módulos que forman FOG, es necesario comentar un componente ajeno al flujo de datos. Este componente es el módulo de configuración (CONFIG, mostrado en la figura 1), en el cual el usuario de la herramienta establece los parámetros que determinan la estrategia a seguir para la creación de la ontología. El módulo de configuración lee los datos de un archivo XML, de forma que mediante reflexión se procede a la carga de los mismos definiendo así el flujo de ejecución del sistema.

A lo largo de un proceso intervienen distintos módulos que pueden actuar en tres fases principales:

- *Extracción de datos:* En esta fase se procede a la extracción de los datos desde la fuente indicada en el módulo de configuración. La extracción de datos realiza la indexación de todos los términos que contiene la colección a modelar. Esta colección puede ser una colección de documentos, una base de datos, etc. A lo largo de esta fase intervienen dos componentes principales: el componente básico de indexación de datos y el módulo de preprocesamiento de datos.
- *Modelado de información:* En esta fase se hace uso del concepto entidad. Se entiende por entidad, a la unidad base en la estructura final de FOG. Una entidad se puede definir como un elemento de la colección de documentos a procesar junto con las propiedades que posee en el contexto en el cual es analizada.

La elección de que es una entidad, es introducida en el módulo de configuración por el usuario. Más formalmente una entidad queda definida como una dupla $E = [V, P]$ donde: V es el elemento elegido en el módulo de configuración como unidad base y P es el conjunto de propiedades necesarias para el modelado.

Al cargar la información introducida por el usuario en el módulo de configuración, se

detectan las entidades, las relaciones existentes entre ellas y se ejecutan las funcionalidades que caracterizan el modelado final del grafo. Tales funcionalidades pueden ser: la fusión de entidades, la poda de entidades que no tengan relevancia, etc., dando como resultado un grafo que será la representación del dominio sobre el que se trabaja.

- *Representación de la ontología:* Es la fase final a la que llega el flujo de ejecución de FOG. En este módulo pueden ejecutarse distintos visualizadores y se da un formato concreto a la ontología resultante para poder ser utilizada por un sistema de recuperación de información.

En la figura 1 se indica la arquitectura general de FOG que a continuación será comentada.

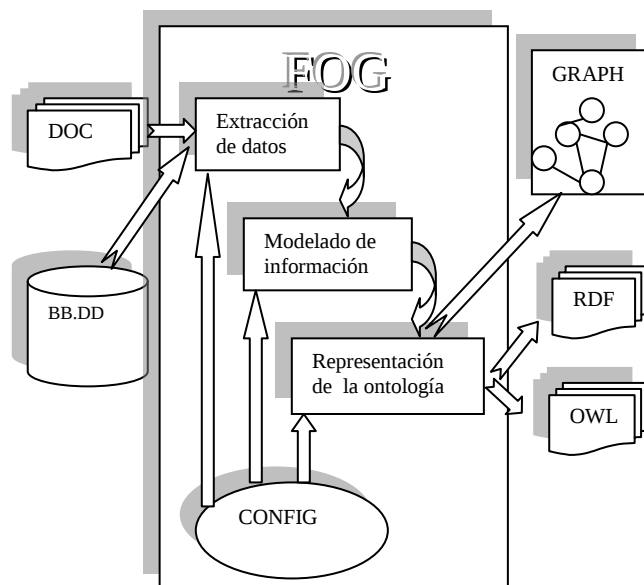


Figura 1: Arquitectura de FOG.

3 EXTRACCIÓN DE DATOS

El módulo de extracción de datos se encarga de extraer los términos contenidos en los documentos de la colección y crear un índice con ellos. Este índice es procesado para poder obtener todos los datos que se encuentran en un documento desde el punto de vista léxico y así, posteriormente, poder darles un valor semántico.

3.1. INDEXACIÓN DE DATOS

Dentro de la fase de extracción de datos, un primer proceso es crear un índice en memoria local con el

² RDF <http://www.w3.org/RDF/>

³ <http://www.w3.org/2004/OWL/>

contenido de la colección, de forma que luego pueda ser utilizado. Con este proceso se consigue independizar la fuente documental de los datos útiles para el proceso de FOG.

3.2. PREPROCESO DE DATOS

El análisis léxico y sintáctico de los datos que contiene el documento se realiza en esta fase. Los datos que se pueden obtener, y que nos interesan para la estrategia de modelado, se obtienen de la configuración introducida por el usuario. También el usuario introduce en el módulo de configuración la herramienta de procesamiento de lenguaje natural, como puede ser GATE⁴ (General Architecture for Text Engineering) entre otras, que permitan operaciones básicas para el procesamiento de lenguaje natural como:

- Tokenizador: analiza un texto en el que identifica palabras, números, símbolos, símbolos de puntuación, espacios en blanco, etc.
- Formateador de oraciones: identifica las oraciones, especificando el inicio y el final de las mismas.
- Etiquetador: identifica para cada palabra el tipo, es decir, indica si es sustantivo, nombre propio, verbo transitivo, adjetivo calificativo, etc.
- Analizador sintáctico: se obtiene un análisis sintáctico del documento a procesar.

4 MODELADO DE LA INFORMACIÓN

En la fase de modelado de la información intervienen distintas funcionalidades necesarias para la generación automática de la ontología. Estas funcionalidades se pueden clasificar entre funcionalidades obligatorias como la funcionalidad de procesado de entidades o detección de relaciones y otras opcionales como la fusión de entidades, la poda de entidades.

FOG es una herramienta que se encuentra en sus inicios, por lo que se puede aumentar el número de funcionalidades opcionales que intervienen en el proceso.

4.1. PROCESADO DE ENTIDADES

El procesado de entidades se basa en un algoritmo capaz de detectar los elementos del documento que inicialmente se consideran entidad. Las características de una entidad

son introducidas por el usuario en el módulo de configuración. Este algoritmo detecta las entidades y extrae del documento toda la información necesaria para la detección de relaciones.

Para la obtención de la información de una entidad se tiene que hacer uso de los datos obtenidos durante el preprocesamiento de datos. Por tanto, alguna de las propiedades que componen una entidad son:

- Valor de la entidad: es el identificador de la entidad, es decir, palabra que forma la entidad.
- Peso de la entidad: este peso se puede definir a nivel de colección. Indicando la relevancia de la entidad en la colección.
- Identificador de oración: identifica a la oración que contiene la entidad.
- Identificador de documento: identifica al documento que contiene la oración en la que se encuentra la entidad.

4.2 DETECCIÓN DE RELACIONES

El siguiente paso en el flujo de ejecución es la detección de las relaciones. En esta fase se relacionan las entidades según la estrategia diseñada por el usuario. En la estrategia diseñada se pueden tener en cuenta entre otros factores, si las relaciones entre entidades se realizan a nivel de oración, a nivel de documento, en función de una distancia entre palabras establecida por el usuario.

Todas las relaciones detectadas tienen un grado de relación semántica. Por ejemplo, si el usuario determina que las relaciones se establecen a nivel de oración y luego a nivel de documento, para todos los documentos de una colección se crearán arcos entre todas las entidades que formen parte de la misma oración y luego se crearán arcos entre todas las oraciones del documento para terminar uniendo los documentos y estructurar jerárquicamente toda la colección.

El siguiente paso es el cálculo del grado de relación semántica que existe en las conexiones detectadas para cada par de entidades. Para realizar dicho cálculo, una de las técnicas propuestas es la lógica borrosa tal y como ya proponen otros autores comentados en el estado del arte.

Por ejemplo, Pereira [7] menciona el cálculo del grado de relación semántica entre términos en base a la coocurrencia de términos:

$$W_{ij} = \frac{N_{ij}}{N_i + N_j - N_{ij}}$$

⁴ <http://gate.ac.uk/>

Donde N_i (N_j) es el número de documentos que contienen a los términos t_i (t_j) y N_{ij} es el número de documentos que contienen el término t_i y el término t_j .

Esta técnica para establecer el grado de las relaciones, otras propuestas en el estado del arte y nuevas técnicas que se realizarán en trabajos futuros, tienen cabida en FOG. El usuario decide que técnica aplicar introduciendo las variaciones en el módulo de configuración.

4.3 FUSIÓN DE ENTIDADES

Es normal que en un documento existan palabras que al compartir una estrecha relación semántica, pertenezcan a un mismo subconjunto del dominio que estamos tratando. Por tanto, en toda una colección documental perteneciente al mismo dominio, el número de elementos que componen este subconjunto aumentan. Es por este hecho que el grafo resultante tras el proceso de detección de relaciones puede llegar a ser ingobernable. De aquí surgió la idea de la fusión. Esta funcionalidad se basa en algoritmos que detectan las entidades que tienen un determinado grado de relación semántica y las agrupa formando una única entidad de mayor nivel de abstracción. Al finalizar este proceso se incorporan al grafo conceptual nuevas entidades catalogadas como conceptos. Se puede definir concepto como el conjunto de entidades que comparten un cierto grado de similitud semántico.

4.4 PODA DE ENTIDADES

Cuando la colección documental alcanza un tamaño considerable y toda la colección pertenece a un mismo dominio, el grado de relevancia a nivel de colección de algunas entidades, en comparación con el grado de relevancia obtenido por otras entidades, se puede encontrar fuera del umbral establecido por el usuario. Este hecho indica que esta entidad es poco significativa para el dominio del que se trata y se descarta para la concepción final del grafo conceptual.

5 REPRESENTACIÓN DE LA ONTOLOGÍA

Tras la conclusión de los procesos anteriores se tiene como resultado una ontología. Pero su utilización en otras aplicaciones depende de la representación final. Para sistemas de recuperación de información como BUDI [8], la representación de un grafo conceptual de un dominio facilitaría la desambiguación de términos especificados en una consulta y así poder precisar más la respuesta del sistema a una consulta de un usuario basándose en el contexto.

FOG permite la visualización de la ontología en distintos formatos y cada uno pone de manifiesto una serie de características a destacar. Como ejemplo de visualización se presentan las siguientes figuras. Estas muestran las relaciones entre las entidades detectadas en los siguientes documentos que son oraciones de la colección MEDLARS⁵:

- Doc1: *correlation between maternal and fetal plasma levels of glucose and free fatty acids.*
- Doc2: *free fatty acid concentration in maternal plasma and fetal body fat content.*
- Doc3: *the concentration of non-esterified fatty acids in maternal and fetal plasma in intact, alloxan-diabetic and x-ray-irradiated rats.*
- Doc4: *7 patients who had biopsy of the bone marrow, all had abnormal types or numbers of plasma cells.*

En la figura 2, se pueden apreciar la expansión de conceptos. El usuario puede seleccionar el nivel de abstracción que desea visualizar.

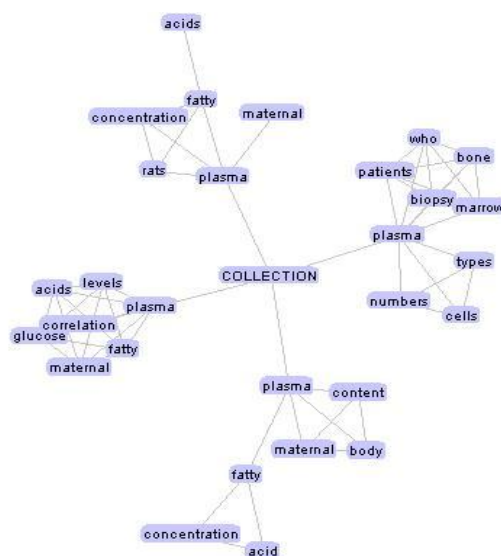


Figura 2: Representación visual de la ontología en función de nivel jerárquico.

En la figura anterior se puede ver la relación existente entre la entidad *acids* y la entidad *fatty*, y por consiguiente, existe un grado de relación semántica establecido por la fórmula aplicada en la etapa de detección de relaciones. Un ejemplo de cálculo del grado de relación semántica puede ser la utilización de la fórmula comentada en el apartado 4.2.

⁵ <ftp://ftp.cs.cornell.edu/pub/smart/>

En la figura 3, se representa la proximidad entre entidades en función de una entidad elegida como entidad central. En el caso del ejemplo, la entidad central es COLLECTION.

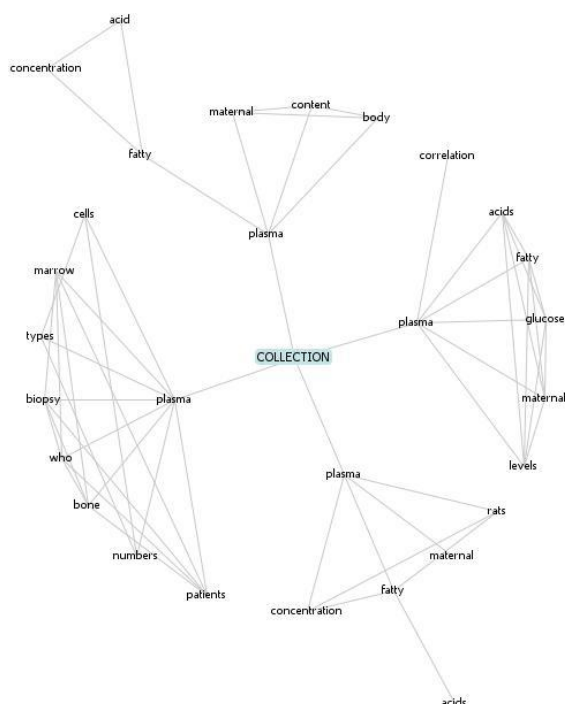


Figura 3: Representación visual de la ontología en función de una entidad central.

Para que un sistema de recuperación de información pueda hacer uso del grafo conceptual, éste debe estar representado en un lenguaje que el sistema de recuperación de información pueda manejar. La representación final del grafo conceptual podría ser formateada en un lenguaje como RDF, OWL ó FUZZY-OWL [2]. Se hará uso de una representación u otra en función de lo especificado por el usuario y de las restricciones que se incluyan dentro de la ontología.

6 CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se ha presentado la arquitectura de la plataforma FOG, la cual está orientada al proceso de generación automática de ontologías en colecciones de datos tales como colecciones documentales, bases de datos relacionales, etc.

Actualmente la plataforma está preparada para la incorporación de algoritmos, así pues el siguiente paso es la incorporación de algoritmos que permitan generar los distintos componentes de una ontología así como incorporar algunas de las propuestas de los autores que se

han citado en el estado del arte para poder comparar sus resultados con propuestas propias.

De igual forma otro de los próximos trabajos será la creación de las interfaces necesarias para conectar la plataforma BUDI con FOG, con el fin de aprovechar sus resultados para distintas fases como la expansión de consultas o la desambiguación de significados en las distintas fases del proceso de metabúsqueda.

Agradecimientos

La investigación realizada está financiada por el proyecto TIN2007-67494 del Ministerio de Educación y Ciencia, el proyecto PEIC-09-0196-3018 de la Junta de Comunidades de Castilla-La Mancha y el proyecto TSI-020302-2009-55 financiado por el Ministerio de Industria, Turismo y Comercio.

Referencias

- [1] Baeza-Yates, R.A., Ribeiro-Neto, B., Modern Information Retrieval, Boston, MA, USA, Addison Wesley, 1999.
- [2] Calegari, S., Ciucci, D. (2007). Fuzzy Ontology and Fuzzy-OWL in the KAON Project. Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International. 1-6.
- [3] Gruber, T. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5 (2), 199-220.
- [4] Horng, Y., Chen, S., & Lee, C. (2003). Automatically constructing multi-relationship fuzzy concept networks for document retrieval. Applied Artificial Intelligence, 17(4), 303-328.
- [5] Liang, T., & Chang, C. (1999). Chinese textual retrieval based on fuzzy concept network. Proceedings of National Computer Symposium, 1, 61-67.
- [6] Lucarella, D., & Morara, R. (1991). FIRST: Fuzzy information retrieval system. Journal of Information Science, 17(1), 81-91.
- [7] Pereira, R., Ricarte, I., & Gomide, F. (2009). Information Retrieval with FROM: The Fuzzy Relational Ontological Model. International Journal, 24, 340-356.
- [8] Serrano-Gurrero, J., Romero, F., Olivas, J., de La Mata, J., & Soto, A. (2009). BUDI: Architecture for Fuzzy Search in Documental Repositories. Mathware and Soft Computing, 16, 71-85.
- [9] Tho, Q., Hui, S., Fong, A., & Cao, T. (2006). Automatic fuzzy ontology generation for semantic web. IEEE Transactions on knowledge and data engineering, 18 (6), 842-856.
- [10] Zadeh, L. (1965). Fuzzy sets. Information and Control, 8, 338-353.