Enfoque Conceptual Borroso en Recuperación de Información

Prof. Dr. Andrés Soto Villaverde
Universidad Autónoma del Carmen
Cd. Carmen, Campeche, México
Email: soto_andres@yahoo.com

Breve reseña

- Doctor en Informática, Universidad de Castilla La Mancha, Ciudad Real, España
- Licenciado en Matemáticas, Universidad de La Habana, Cuba
- Profesor Titular Universidad Autónoma del Carmen, Ciudad del Carmen, Campeche, México
- Ha sido profesor de Computación en Universidad de La Habana, Cuba, y en Universidad Simón Bolívar y Universidad Católica Andrés Bello, de Caracas, Venezuela

Temas de interés

- Clasificación y agrupamiento (clustering) de documentos en lenguaje natural (inglés, español) utilizando bases de conocimientos, ontologías, conjuntos y lógica difusa, etc.
- Sistemas Pregunta Respuesta basados en metabuscadores , así como en el uso de restricciones generalizadas, protoforms y ontologías como YAGO

Otros temas

- También nos interesan temas tales como: Ontologías, Redes semánticas, Extracción de Información, Representación del Conocimiento, Restricciones generalizadas y protoforms
- A los cuales no les hemos podido dedicar el tiempo que desearíamos

Enfoque Difuso

- Se han definido índices difusos (fuzzy) para reflejar relaciones tales como la sinonimia y la relación entre el término definido y los términos utilizados en su definición.
- Objetivo: Priorizar los aspectos de significado conceptual (semánticos) de los términos, más que los aspectos léxicos o estadísticos

Grado de sinonimia

- El grado de sinonimia entre dos términos se expresa mediante:

$$S(t_1, t_2) = \frac{|M(t_1) \cap M(t_2)|}{|M(t_1)|}$$

- Donde *M(t)* representa el conjunto de significados del término t
- Ejemplos *S*(*auto*, *automobile*)=1, *S*(*automobile*, *auto*)=0.5
- Observar que la relación no es simétrica

Autor: Andrés Soto
soto_andres@yahoo.com

Grado de polisemia

- El grado de polisemia de un término *t* expresa la debilidad del mismo:
  - más interpretaciones => más polisemico, más debil;
  - menos interpretaciones => menos polisémico, más fuerte.

$$I_P(t) = 1 - \frac{1}{N_m(t)}$$

- $N_m(t)$: número de significados de *t*

Autor: Andrés Soto
soto_andres@yahoo.com

Frecuencia conceptual

- El índice de frecuencia conceptual expresa la frecuencia con que se hace referencia a un concepto *m* en el documento D*j*

$$Cf_j(m) = \frac{\sum_{t_i \in T(m)} \frac{n_{ij}}{N_m(t)}}{n_{*j}}$$

- n_{ij}: # de veces que aparece el término t_i en j y n_{*j} # de términos contenidos en j

Autor: Andrés Soto
soto_andres@yahoo.com

Significado de una palabra

- Los diccionarios recogen y explican el significado o definición de las palabras o términos.
- La definición expresa el uso, la función y la esencia del concepto asociado al término
- Las palabras que forman parte de la definición de un término dado tienen una relación estrecha con el concepto o significado que explican

Medir la presencia de un concepto dado

- La expresión siguiente busca medir la presencia de un concepto dado en un documento a través de medir la presencia de los términos que forman parte de su definición

$$Def_j(m) = \frac{\sum_{t_i \in B(m)} n_{i,j}^2}{n_{*,j}}$$

Autor: Andrés Soto
soto_andres@yahoo.com

Medida de similitud

- La medida de similitud del coseno entre dos vectores D1 y D2 nos permite medir la similitud entre los índices de frecuencia conceptual de dos documentos

$$\text{Cos} = \frac{\vec{X} \cdot \vec{Y}}{\left\|\vec{X}\right\|_2 \left\|\vec{Y}\right\|_2}$$

APLICACIONES DE ESTOS MODELOS EN CONJUNCIÓN CON MÉTODOS DE AGRUPAMIENTO

Autor: Andrés Soto
soto_andres@yahoo.com

Modelo de Agrupamiento Híbrido (paso 1)

- El método de agrupamiento híbrido combina dos métodos de agrupamiento diferentes.
- Se utiliza un método de agrupamiento jerárquico para detectar las relaciones entre documentos.
- Cada documento define inicialmente su propio grupo (cluster).

Etapa de agrupamiento

- Posteriormente, los clusters se van integrando por parejas sucesivamente hasta que queda formado el árbol completo utilizando las relaciones de similitud antes explicadas y una técnica de enrejado modificada.

Fin del paso 1

- El proceso de agrupamiento inicial termina cuando los documentos están agrupados en conjuntos que corresponden a conceptos diferentes con determinado umbral de similitud entre los documentos de un mismo grupo.

Paso 2

- En el segundo paso se lleva a cabo un proceso iterativo donde cada documento es colocado en los clusters cuyo promedio de similitud es mayor que el umbral de similitud utilizado previamente.
- El algoritmo incluye mezclar y eliminar clusters cuyo umbral sea bajo i.e. los clusters con demasiados elementos o con poca similitud entre ellos.

Resultados obtenidos

comparación con métodos de agrupamiento basados en tf-idf y fuzzy c-means con las colecciones SMART y REUTERS utilizando varias métricas

	TF-IDF & FCM	Hybrid Model	TF-IDF & FCM	Hybrid Model
Metric	SMART	SMART	REUTER	REUTER
MS	37	49	29	45
CSS	24	55	22	43
F-measure	43	63	45	54
NO	22	10	25	15
SNC	15	8	28	10

Mapas de Kohonen

- Los mapas auto organizados (Self-Organizing Maps o mapas de Kohonen) son redes neuronales que representan el espacio de entrada utilizando una estructura topológica en una rejilla que almacena las relaciones de vecindad.

Agrupamiento con SOM

- Su mayor ventaja es el uso de aprendizaje no supervisado para descubrir la estructura subyacente de los datos en las tareas de clustering.
- La función de similitud entre documentos y neuronas se definió en base al modelo de similitud conceptual entre vectores visto anteriormente y el coeficiente de Jaccard

Resultados SOM++

En comparación con otros algoritmos de clustering como el SOM básico o mejorado, el BBK (Bisecting k-means using background knowledge) y el CFWMS (Clustering based on Frequent Word Meaning Sequences)

Algorithm	*F*-measure
Basic SOM	0.53
SOM+	0.46
CFWS	0.57
BBK	0.46
SOM++	0.76

CLASIFICACIÓN BASADA EN ECUACIONES DE PROXIMIDAD Y ONTOLOGÍAS

Autor: Andrés Soto
soto_andres@yahoo.com

Enfoque declarativo difuso

- Actualmente, el enorme crecimiento de la WWW, hace más costoso o difícil disponer de conjuntos de documentos correctamente etiquetados para entrenar los algoritmos de aprendizaje automático para clasificación de documentos.
- Por ello resulta conveniente evaluar otros métodos que no requieran conjuntos de documentos previamente etiquetados.

Ecuaciones de proximidad

- Uno de estos enfoques consiste en categorizar textos no etiquetados utilizando solamente una ontología de términos modelada mediante un conjunto de ecuaciones de proximidad obtenidas a partir de diferentes ontologías y tesauros utiliza una extensión difusa (fuzzy) de Prolog

Resultados obtenidos

- El experimento consistió en clasificar 173 artículos ya clasificados de la colección Reuters 21578 correspondientes a 6 categorías no excluyentes.

Relación de proximidad	Correcto	Mal clasificado	No clasificado
Wikipedia/Vector	81%	13%	6%
WordNet/Vector	73%	9%	17%
Contexto	65%	3%	32%
Sin ontología	10%	2%	88%

Definición conceptual de las categorías

- Otra metodología consiste en clasificar documentos a partir de la definición conceptual de las categorías.
- Se basa en construir una base de conocimientos alrededor de cada categoría, empleando conjuntos y relaciones borrosas, que permita asociarles el máximo contenido semántico posible.

Enfoque basado en reglas

- A diferencia de los enfoques basados en reglas de la década de los 80s, la base de conocimientos se construye a partir de relaciones tomadas de diccionarios y tesauros como WordNet y EUROVOC, entre otros.

Experimentos

1. Experimento E1 para comprobar en cuántos documentos aparecía referenciada una categoría de Reuters midiendo solamente la presencia del término
2. Experimento E2 donde ya se incluyen términos sinónimos de dicha categoria o algún otro término relacionado (i.e. earn – earnings).

Resultados obtenidos

	Precision		Recall		F1	
Tópico	**E1**	**E2**	**E1**	**E2**	**E1**	**E2**
'earn'	0,37	0,37	0,03	0,89	0,05	0,52
'crude'	0,69	0,29	0,47	0,97	0,56	0,45
'grain'	0,68	0,67	0,44	0,53	0,54	0,59
'trade'	0,22	0,17	0,93	0,93	0,35	0,28
'interest'	0,10	0,07	0,39	0,39	0,15	0,12
'wheat'	0,65	0,64	0,97	0,97	0,78	0,78
'ship'	0,48	0,24	0,51	0,58	0,49	0,34
'corn'	0,60	0,64	0,73	1,00	0,66	0,78

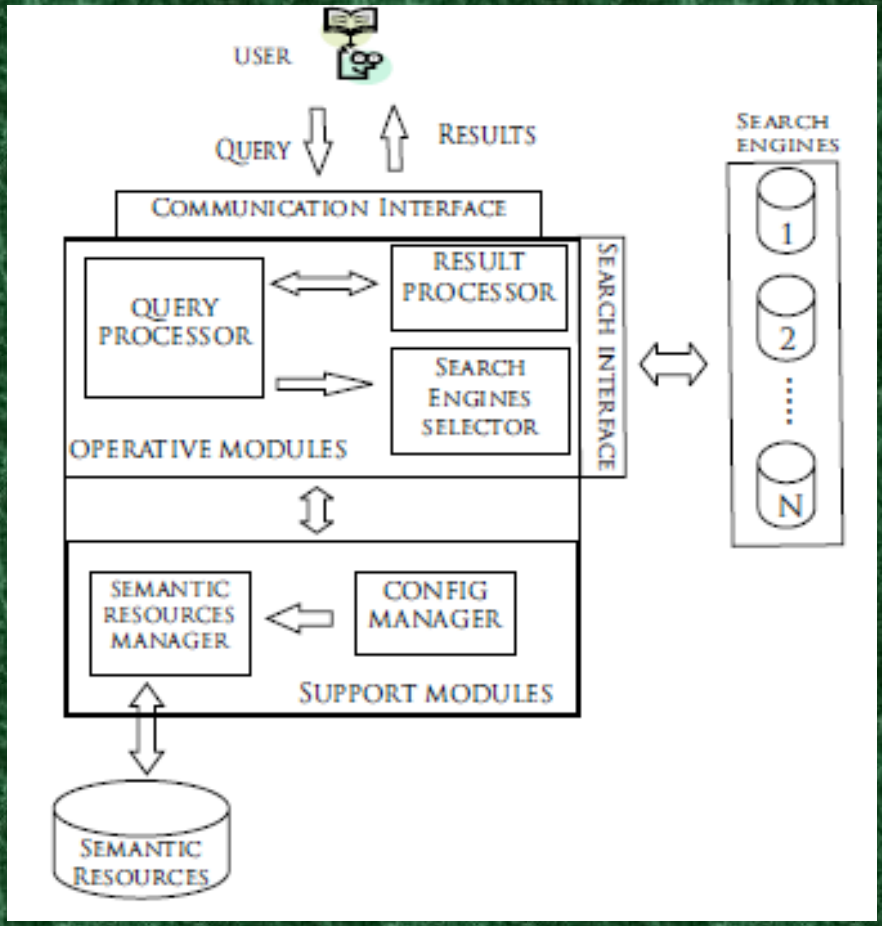
Question Answering Systems

- Por otra parte se trabaja en sistemas de pregunta respuesta basados en metabuscadores como Castalia o QAS YAGO basado en la ontología YAGO.
- También se han utilizado metabuscadores como BUDI para recuperar oraciones relacionadas con una pregunta dada, las cuales son procesadas posteriormente.

Sistema BUDI

- BUDI es un metabuscador especialmente diseñado para utilizar operaciones difusas.
- Su arquitectura, en especial la estructura de representacion de las consultas, está diseñada para trabajar con operadores difusos como t-normas, t-conormas, etc, las que se pueden emplear para la expansión de las consultas, agrupamiento, etc

Arquitectura de BUDI



Sistema IRKA

- El metabuscador BUDI se utiliza aquí para recopilar los documentos relacionados con una pregunta dada.
- Luego, el sistema IRKA se extraen aquellas oraciones que contienen términos incluidos en la pregunta: mientras más términos de la pregunta aparezcan en la oración, mayor será la relación de la misma con la respuesta

Base de Conocimientos

- Las oraciones recopiladas de los diferentes documentos se integran en una base de conocimientos formada por oraciones copulativas, comparativas y superlativas con información relacionada con la pregunta realizada junto con otras relaciones obtenidas a partir de diferentes ontologías.

Inferencia

- La pregunta original entonces se transforma en un objetivo que se trata de demostrar (inferir) utilizando dicha base de conocimientos, aplicando reglas de razonamiento basadas en la lógica difusa y la teoría de las protoformas de Zadeh

Arquitectura de IRKA

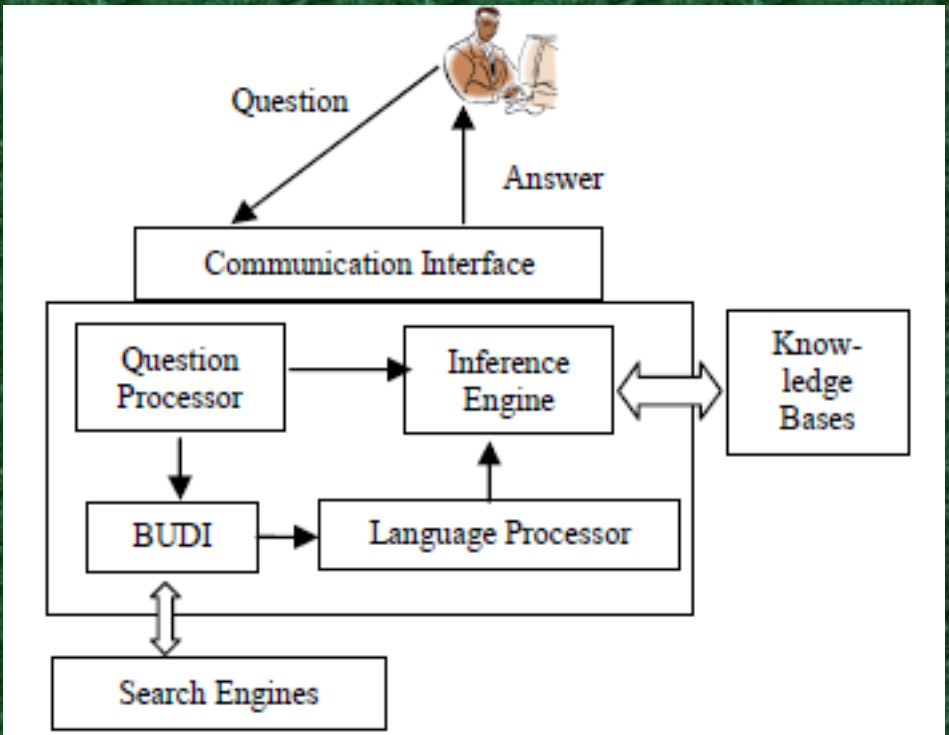


Fig. 1 IRKA architecture

Conclusiones

- Hemos querido mostrar las áreas y temas de interés de nuestras investigaciones, así como los principales resultados obtenidos hasta el momento para darnos a conocer.
- Gran parte de este trabajo se ha realizado en colaboración con el grupo SMILe (Soft Management of Internet and Learning) de la Universidad de Castilla La Mancha

Temas de interés

- Como se desprende de la presentación, nos interesan temas tales como Recuperación de Información, Sistemas Pregunta Respuesta (Question Answering), Ontologías, Redes semánticas, Extracción de Información, Representación del Conocimiento, Restricciones generalizadas y protoforms

Interes por colaborar

- Quiero resaltar mi interés personal por colaborar con colegas de otras instituciones, dado que en la universidad donde trabajo no tengo otros colaboradores, ni estudiantes.
- También conspira en mi contra la lejanía geográfica con relación a otras instituciones de México con más posibilidades de desarrollo.

¡MUCHAS GRACIAS!

PREGUNTAS?

Autor: Andrés Soto
soto_andres@yahoo.com