Creado el 21/01/2014
Última versión: 21/01/2014
Miguel Ángel García Cumbreras, Universidad de Jaén
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Aplicaciones de las TLH/RECUPERACIÓN DE INFORMACIÓN

BÚSQUEDA DE RESPUESTAS

Dentro de la recuperación de información se identifica una rama especializada en responder, normalmente de forma breve, a preguntas concretas. Mientras la recuperación de información en general hace eso, encontrar información que después debemos procesar y filtrar nosotros mismos, muy habitualmente páginas web, la búsqueda de respuestas intenta extraer de ese conjunto de información datos específicos que satisfagan la pregunta concreta que hemos lanzado: ¿dónde está el museo del Louvre?

Transcripción del vídeo

Hola, soy Miguel Ángel García, profesor de la Universidad de Jaén y miembro del grupo de investigación SINAI. Te quiero mostrar el tema de la Búsqueda de Respuestas, correspondiente a la unidad de Recuperación de Información.

Los objetivos de esta lección son los siguientes:

- En primer lugar, definiremos y motivaremos el problema de la búsqueda de respuestas.
- A continuación, repasaremos los distintos módulos que componen este tipo de sistemas.
- Finalmente, veremos qué criterios y métricas podemos utilizar para evaluar estos sistemas.

Comenzaremos definiendo la búsqueda de respuestas como una tarea automática realizada por ordenadores, que tiene como finalidad encontrar respuestas concretas a necesidades de información precisas que los usuarios formulan, a preguntas concretas.

La siguiente pregunta que se te puede plantear es: ¿y son necesarios actualmente estos sistemas?

Sin entrar en muchos detalles históricos, el origen de los sistemas de búsqueda de respuestas se remonta al año 1961, como una de las tareas más antiguas de PLN, y que Simmons y otros colegas relataron en 1964. Pero la justificación más racional es que la gran cantidad de información no estructurada y la necesidad de encontrar respuestas precisas de forma más rápida, provoca el gran desarrollo en los últimos 10 años de este tipo de sistemas.

En esta diapositiva vemos un ejemplo con una pregunta sencilla, ¿quién escribió El Quijote?. En nuestra colección documental podemos encontrar diversas respuestas potenciales, y manualmente buscaríamos algún documento donde aparezcan al menos los términos “El Quijote” y escribir o algún término relacionado con este verbo, tal como relatar, escritor, autor, etc.

Echando un vistazo en las posibles respuestas podemos adivinar que en algunos casos obtener una respuesta correcta no parece complejo, pero en otros sí que es necesario tener conocimiento externo o leer varias veces la frase para obtener la respuesta correcta. Por este motivo, los sistemas de búsqueda de respuesta admiten en sus procesos internos una gran cantidad de recursos y aplicaciones de pln (detección de entidades, anáfora y correferencia, etc.)

Continuaremos con los tipos de preguntas que de forma más habitual un los sistemas de búsqueda de respuesta abarcan. En primer lugar encontramos las preguntas factuales, que son aquellas sobre hechos concretos, una fecha, una persona, un lugar, un valor numérico. Son más fáciles de tratar para estos sistemas y algunos ejemplos son:

- ¿Quién escribió «El Quijote»?
- ¿Cuántas calorías tiene un plátano?
- ¿Qué porcentaje de españoles es vegetariano?
- ¿Cuál es la ciudad con más playas?

También se intentan resolver cuestiones más complejas, narrativas, que requieren de conocimiento externo y de una redacción más compleja a nivel de producción de lenguaje, para dar una respuesta concreta. Algunos ejemplos de este segundo grupo son:

- En personas con dislexia, ¿cómo afecta el aprendizaje de idiomas?
- ¿Qué opinan los estudiantes de bachillerato del actual modelo educativo?

Otro tipo de preguntas actualmente abordado por sistemas automáticos, son las preguntas de definición, preguntas sobre cosas concretas cuya respuesta es una definición. En este caso particular la longitud de la respuesta es un punto débil, ya que no es fácil identificar cuál es el texto mínimo que es una respuesta para el usuario. Por ejemplo, en este grupo están las preguntas:

- ¿Qué es BMW?
- ¿Qué es una bacteria?

Los sistemas comerciales, que tienen la restricción del tiempo real, trabajan normalmente con preguntas factuales sencillas del tipo:

- ¿Dónde está el museo del Prado? Con la respuesta correcta: En Madrid (España)
- ¿Cuál es la abreviatura de Compact Disc? Con la respuesta correcta: CD
- ¿Qué frutos secos tienen cáscara y fruto verdoso? Con la respuesta correcta: los pistachos
- ¿Cuál es el teléfono de emergencias? Con la respuesta correcta: El 112

Encontramos dos paradigmas a la hora de desarrollar un sistema de búsqueda de respuestas. Por un lado, aquellos sistemas basados en recuperación de información, y que suelen aplicar diversas tecnologías del lenguaje. En este grupo encontramos los sistemas desarrollados y presentados en las conferencias internacionales TREC y CLEF, o Google.

Por otro lado encontramos los sistemas basados en bases de conocimiento, normalmente en datos de una base de datos, y sistemas híbridos, que no requieren ningún procesamiento con tecnologías del lenguaje para la extracción de las respuestas. En este segundo grupo están los sistemas Siri de Apple, el de Wolfram Alpha o algunos de los asistentes virtuales más utilizados, como el de Ikea.

Para algunas preguntas sencillas, basta con buscarlas en algún buscador web como el de Google, para encontrar fácilmente la respuesta. Por ejemplo, si ponemos en Google: ¿Quién escribió el Quijote?, encontramos rápidamente la respuesta dado que Google, a partir de toda la información

textual que tiene indexada es capaz de indicarnos la respuesta concreta.

Y si la pregunta contiene un entidad muy concreta, esa misma entidad hará que en buscadores web podamos encontrar la respuesta buscada, como en el ejemplo ¿dónde podemos encontrar el museo del prado?, aunque está vez tengamos que ser nosotros quienes saquemos la respuesta concreta, ya que los buscadores web no nos la van a indicar.

En esta diapositiva vemos el esquema básico de un sistema de búsqueda de respuestas para preguntas factuales, basado en recuperación de información.

Destacamos 3 módulos fundamentales, el de procesado de la pregunta, el de recuperación de información y el de extracción de respuestas.

Ante la entrada de una pregunta, un primer módulo fundamental es el del procesado de la misma. Además de otras operaciones como la formulación de la consulta un paso fundamental es la detección del tipo de pregunta, para conocer el tipo de respuesta esperado. Por ejemplo, ante la pregunta ¿quién descubrió América?, es fundamental en este paso saber que el tipo de pregunta es sobre una persona, de modo que entre las posibles respuestas tendremos que descartar lo que no sean nombres de personas.

En el segundo módulo partimos de una colección documental que tenemos que indexar, utilizando algún sistema de recuperación de información, y que una vez lanzada la consulta formulada en el primer paso, nos devolverá un conjunto de documentos ordenados por relevancia. A continuación, y con el fin de simplificar la tercera fase, sobre los documentos relevantes se aplica una fase de selección de pasajes relevantes, aquellos que se presupone que contienen las respuestas concretas.

Existe otra opción muy utilizada por los sistemas de búsqueda de respuesta, y consiste en partir cada documento de la colección en pasajes de X frases, y crear un índice con estos pasajes. De esta forma evitamos la etapa de selección de pasajes ya que el sistema de recuperación de información devolverá directamente pasajes relevantes. Además se van a devolver pasajes relevantes de documentos que no son demasiado relevantes, que con la primera opción no se devolverían, por lo que mejoramos el rendimiento de este módulo.

En la tercera fase, extracción de respuestas, realizamos un primer paso de selección de respuestas candidatas, aplicando distintos métodos. Estas respuestas candidatas son puntuadas para finalmente devolver un conjunto de posibles respuestas al usuario. En esta fase podemos hacer uso, por ejemplo, de evidencias del texto, usaremos el tipo de respuesta esperado y también podemos utilizar recursos externos.

En cuanto a los sistemas basados en conocimiento, tal como Siri o Wolfram Alpha, identificamos dos procesos principales: el primero de representación semántica de la pregunta y el segundo de mapeo de esta representación a una consulta estructurada.

En el primer paso obtenemos el campo semántico de aplicación de la consulta: si es una entidad, una fecha, una hora, si pregunta por una persona, por una localización, por una cantidad numérica, etc.

En el mapeado pasamos esa representación semántica concreta a una consulta estructurada o lo buscamos en recursos externos, tales como:

- bases de datos geoespaciales, para consultas georeferenciadas

- ontologías (Wikipedia, DBpedia, Wordnet)
- Sistemas de reservas de viajes, o de opiniones de restaurantes
- Bases de datos científicas, etc.

Y de esa forma obtendremos las respuestas concretas.

Los sistemas híbridos combinan ambas cosas y de esta forma aplican, en un primer paso, métodos de procesado de la pregunta creando una representación pseudosemántica y, en un segundo paso, aplicamos recuperación de información, para extraer respuestas candidatas. A estas respuestas encontradas en la colección documental, le podemos añadir otras obtenidas de recursos externos y de datos estructurados.

Finalmente, las respuestas candidatas son puntuadas utilizando diversos recursos basados en conocimiento, tales como base de datos geoespaciales de nuevo, razonamiento temporal o alguna taxonomía.

En cuanto a la evaluación de este tipo de sistemas, nos vamos a centrar en los que hacen uso de Recuperación de Información.

En este caso, las métricas comunes son las mismas que en Recuperación de información, la precisión y el recall, y también son aplicables en los sistemas de búsqueda de repuestas, pero es más conveniente aplicar la medida llamada Mean Reciprocal Rank o MRR. Es una medida creada para este tipo de sistemas que tiene en cuenta tanto si la respuesta es correcta o no, como la posición de esa respuesta en la lista de posibles respuestas que cada sistema devuelve. Es este caso es más adecuada porque se considera muy importante la posición de la respuesta correcta.

Además de esta medida general, los sistemas de búsqueda de respuestas se pueden medir en varios puntos. Se puede evaluar la bondad del clasificador de preguntas, en la primera etapa de procesado de la pregunta, el rendimiento del sistema de recuperación de información, y también esta última fase por separado de extracción de respuestas. Hay que tener en cuenta y destacar que un mal funcionamiento de los primeros módulos provocará que, en cualquier caso, se obtenga un mal rendimiento del sistema, por lo que es importante optimizar cada módulo por separado.

Esta evaluación se puede realizar de forma automática, aunque es bastante complejo. Necesitaríamos trabajar con un entorno controlado, formado por una colección documental, preguntas ya preparadas con distintos niveles de complejidad y todas las respuestas correctas para cada una de estas preguntas. Y esas respuestas las tendríamos que sacar de la colección de forma manual para intentar no dejarnos ninguna fuera.

En cuanto a las respuestas pueden ser correctas pero no exactas si, por ejemplo, no son la cadena de texto mínima que el usuario puede considerar como una respuesta esperada. Por ejemplo, ante la pregunta “¿qué es BMW?” una respuesta correcta sería “Una marca de coches alemana”, pero no exacta sería “Una marca de coches alemana, que anualmente factura una cantidad de 200 millones de euros y cuya sede social se encuentra en Berlín”.

Para terminar con la lección, daremos un repaso a lo aprendido.

De nuevo definir los sistemas de búsqueda de respuestas como aquellos que de forma automática devuelven respuestas concretas a preguntas precisas de los usuarios.

Destacamos los dos paradigmas, los sistemas basados en recuperación de información y los basados en conocimiento y fuentes estructuradas, y destacamos los tres tipos de preguntas que normalmente se abordan: las factuales, las narrativas y las de definición.

Con esto finalizamos este apartado del curso.

Puedes encontrar más información sobre el curso en su página web y plantear tus dudas y comentarios a través de correo electrónico o a través de Twitter.

Muchas gracias por la atención.