

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 638 576**

51 Int. Cl.:

G06F 17/27 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **15.04.2008 PCT/EP2008/054558**

87 Fecha y número de publicación internacional: **23.10.2008 WO08125676**

96 Fecha de presentación y número de la solicitud europea: **15.04.2008 E 08736245 (5)**

97 Fecha y número de publicación de la concesión europea: **07.06.2017 EP 2137638**

54 Título: **Un método para la extracción de patrones de relación a partir de artículos**

30 Prioridad:

16.04.2007 EP 07106252

45 Fecha de publicación y mención en BOPI de la traducción de la patente:

23.10.2017

73 Titular/es:

**THE EUROPEAN UNION, REPRESENTED BY THE
EUROPEAN COMMISSION (100.0%)
RUE DE LA LOI, 200
1049 BRUSSELS, BE**

72 Inventor/es:

TANEV, HRISTO TANEV

74 Agente/Representante:

ELZABURU, S.L.P

ES 2 638 576 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Un método para la extracción de patrones de relación a partir de artículos

Campo técnico

5 La presente invención se relaciona generalmente al campo técnico del procesamiento del lenguaje natural y más específicamente a un método y sistema para la extracción de patrones de relación a partir de documentos de lenguaje natural.

Antecedentes

10 Expresar una cosa en otras palabras, o “parafrasear”, juega un papel importante en la variedad y complejidad de los documentos de lenguaje natural. Uno puede expresar un único evento en miles de formas en frases de lenguaje natural. Un escritor creativo usa muchas paráfrasis para decir un solo hecho. Esto aumenta enormemente la dificultad del procesamiento del lenguaje natural.

15 El modelado de esta variabilidad semántica en el lenguaje ha atraído mucha atención en los últimos años. Muchas aplicaciones en lenguaje natural, como la Recuperación de Información, la Traducción Automática, la Respuesta a Preguntas, el Resumen de Textos, o la Extracción de Información, tienen que reconocer que el mismo significado puede expresarse en el texto en una enorme variedad de formas superficiales. Se ha dedicado una sustancial investigación a la adquisición de patrones de paráfrasis, que representan varias formas en las que se puede expresar un cierto significado.

20 Además del patrón parafraseado, existen relaciones, que no son exactamente paráfrasis, pero están sin embargo relacionadas y son potencialmente útiles para los sistemas de recuperación de información. Por ejemplo “X adquirió Y” implica “X posee Y”. Resulta que una noción algo más general necesaria para muchas aplicaciones es la de “relaciones de implicación”. Estas “relaciones de implicación” (que incluye patrón de paráfrasis) son relaciones direccionales entre dos expresiones, donde el significado de una se puede implicar a partir del significado de la otra. Estas relaciones proporcionan un amplio marco para representar y reconocer variabilidades semánticas, como se propone en (Dagan and Glickman, 2004). Por ejemplo, si un sistema de Pregunta y Respuesta tiene la respuesta a la pregunta “¿Quién posee YouTube®?” y el cuerpo incluye la frase “Google® adquirió YouTube®”, el sistema puede usar la relación de implicación conocida para concluir que esta frase realmente indica la respuesta deseada.

25 Para realizar tales conclusiones a gran escala, las aplicaciones necesitan poseer una gran base de conocimiento (KB) de patrones de implicaciones. Tradicionalmente, las bases de conocimiento que contienen tales relaciones de implicación se crean manualmente. Sin embargo, esta tarea de ingeniería del conocimiento es inherentemente difícil dado que los humanos no son buenos en la generación de listas completas de reglas.

30 Han sido descritos varios métodos en el pasado para aprender relaciones que usan métodos de núcleo, por ejemplo por Zelenko Dmitry, Aone Chinatsu, Richardella Anthony “Kernel Methods for Relation Extraction”, Journal of machine learning reserach, 2003, vol. 3, no6, pp. 1083-1106 o por A. Culotta, J. Sorensen, “Dependency tree kernel for relation extraction”, Proceedings of the 42nd Annual Meeting of the ACL, 2004”. Su desventaja común es la necesidad de anotar manualmente los datos entrenados y la detección de relación lenta (cada par de las entidades se comprueba). Otra desventaja importante es que la salida del algoritmo de aprendizaje automático es muy difícil, si no imposible de ser entendida y manipulada por un experto humano.

35 Barzilay y Lee (Barzilay, Lee, “Learning to Paraphrase: an unsupervised approach using multiple sequence alignment”, Proceedings of HLT-NAACL 2003) presentan un enfoque de patrón lineal sin supervisión, que se basa en frases alineadas. Dado que no es fácil encontrar muchas frases similares, la cobertura de este enfoque tiene limitaciones. Otro enfoque basado en Web de aprendizaje de patrones sin supervisión se describe en Szpektor, Tanev, Dagan, Coppola, “Scaling Web-based Acquisition of Entailment Relations”, Proceedings of EMNLP 2004. El uso de la web garantiza buena cobertura, sin embargo la velocidad está limitada por la velocidad de los motores de búsqueda públicos.

45 Problema técnico

En consecuencia es un objeto de la presente invención proporcionar un método mejorado para la extracción de patrones de relación a partir de artículos.

Descripción general de la invención

50 Para superar el problema mencionado anteriormente, la presente invención propone un método para formar relaciones de implicación según la reivindicación 1.

El método propuesto hace uso de un procedimiento automático o semiautomático para aprender de manera iterativa patrones a partir de relaciones específicas usando grupos de artículos similares. Para cada relación considerada el usuario proporciona al sistema uno o varios patrones pivote o patrones de entrada, y el sistema devuelve patrones, que expresan el mismo significado de formas diferentes. La principal ventaja del método propuesto yace en los

hechos de que:

- los patrones son más rápidos que los enfoques basados en los métodos de núcleo
- el aprendizaje de patrones a partir de grupos de artículos es más rápido que enfoques similares de aprendizaje de patrones que usan la Web, esto hace factible realizar muchas iteraciones de aprendizaje que mejoran la cobertura

5 - usar grupos de artículos trae mayor precisión. De hecho, las expresiones consideradas como paráfrasis son diferentes de un dominio a otro. Aun cuando dos expresiones pueden ser consideradas como con el mismo significado en un cierto dominio, no es posible generalizarlas a otros dominios. En el presente método, el patrón de entrada se hace corresponder con artículos de otro grupo, esto es con artículos que tienen una temática similar y así están relacionados con el mismo dominio. Resulta que las reglas de implicación, que son identificadas por el método se relacionan con un dominio específico.

- el usuario puede definir la relación en la cual él o ella está interesado.

En una primera realización de la invención, el patrón (p) de entrada es un patrón lineal. En otra realización, el patrón (p) de entrada es un patrón sintáctico y dichos artículos son analizados sintácticamente antes del procesamiento de dichos artículos con respecto al patrón (p) de entrada en c).

15 El método propuesto acepta en sus entradas uno o más patrones P sintácticos o lineales con N ranuras (N>1) que expresan una relación R semántica específica. Por ejemplo, dos patrones sintácticos para la relación "criticar" pueden ser:

PERSONA₁ ← sujeto – criticó – objeto → PERSONA₂

PERSONA₁ ← sujeto – es_un_oponente – de → PERSONA₂

20 Los patrones sintácticos son árboles sintácticos que tienen dos o más ranuras. Cada ranura se puede llenar con un tipo de entidad específico. En el ejemplo anterior, las ranuras están designadas con PERSONA₁ y PERSONA₂; se pueden llenar simplemente por nombres del tipo persona. Las plantillas sintácticas se hacen corresponder contra árboles sintácticos, así no son dependientes del orden de palabra lineal y las palabras adicionales, que pueden aparecer en medio. Por ejemplo, el primer patrón hará corresponder frases como "John Adams, presidente de ABC Corp., criticó agudamente a su colega Adam Howard."

El algoritmo puede aceptar patrones lineales como

PERSONA₁ criticó a PERSONA₂.

30 Diferente a los patrones sintácticos, los lineales dependen del orden de las palabras y las palabras por medio. Por ejemplo, el patrón lineal anterior hace corresponder solo frases como "John Adams criticó a Adam Howard", pero no hará corresponder "John Adams criticó agudamente a Adam Howard".

35 En una posible variante del método, se proporciona un conjunto P completo de patrones (p) de entrada con N ranuras de patrón (N>1) en el paso a), y el patrón extraído se añade al conjunto (P) de patrones. En este caso, el conjunto de patrones (P) forma una especie de base de conocimiento que contiene una pluralidad de patrones diferentes, que expresan un hecho específico de maneras diferentes. Igualmente un conjunto (C) de grupos (c) de artículos (a) puede proporcionarse como entrada, donde cada uno de dichos grupos (c) contiene uno o más artículos relativos a un tema principal común.

40 Dependiendo del número de artículos en el grupo de interés y de la calidad deseada de la extracción de patrones, el método para formar relaciones de implicación puede usar solo una parte del grupo para identificar apariciones de los rellenos de ranuras los artículos o el grupo completo. Si por ejemplo un grupo contiene un alto número de artículos, puede ser suficiente identificar la ocurrencia de los rellenos de ranuras N-tupla solo en una parte de los artículos y sin embargo acabar con una extracción de patrones de alta calidad. En otra realización del método, la aparición de los rellenos de ranuras (s₁, s₂, ..., s_N) se identifica a lo largo de todos los artículos del grupo (c) en ii). Esto asegura que la cobertura de la extracción de relación se maximiza.

45 Se apreciará, que en iii) cada relleno de ranura o entidad es sustituida preferiblemente por una variable de ranura para generar patrones que generalmente abarcan una pluralidad de rellenos de ranuras. Además antes de iv), cada patrón extraído preferiblemente es pesado con respecto al número de frases y el número de N-tuplas de relleno de ranura que soportan el patrón extraído respectivo, y solo esos patrones extraídos, para los cuales el peso excede un umbral predeterminado, son además considerados en iv). Esta selección de patrones con alto peso aumenta la calidad de la extracción de relación de implicación en que solo aquellos patrones son sacados y añadidos a la base de conocimiento, los que probablemente indican el tema más que ocurrencias accidentales del patrón p.

50 Se apreciará, que el método de la presente invención puede operar totalmente automáticamente y sin supervisión. Por otro lado, existe la posibilidad de intervención humana en el proceso de aprendizaje. De hecho, en una variante del método, un experto humano puede evaluar la oportunidad de los patrones recién extraídos y/o añadidos.

Se apreciará además, que la presente descripción también se relaciona con un método para construir una base de conocimiento que contiene relaciones de implicación, que comprende los pasos de extraer relaciones de implicación a partir de un cuerpo (artículos) según el método como se ha descrito anteriormente; y almacenar dicho patrón extraído junto con dicho patrón de entrada como relación de implicación en dicha base de conocimiento. Tal método para construir una base de conocimiento que contiene relaciones de implicación, por lo tanto comprende los pasos de:

- 5 a) proporcionar al menos un patrón (p) de entrada con N ranuras de patrón (N>1), dicho patrón (p) de entrada expresa una relación semántica específica entre N entidades que rellenan las N ranuras de patrón del patrón (p) de entrada como rellenos de ranuras,
- 10 b) proporcionar al menos un grupo (c) de artículos, dichos artículos de dicho grupo (c) relacionados con un tema principal común;
- c) procesar dichos artículos con respecto al patrón (p) de entrada e identificar las identidades que hacen corresponder el tipo semántico de las N ranuras de patrón;
- 15 d) si dicho al menos un patrón de entrada se corresponde con una porción de un artículo (a) de dicho al menos un grupo (c):
 - i) almacenar los N rellenos de ranura (s_1, s_2, \dots, s_N) que se corresponden con las ranuras del patrón (p), y un grupo identificador I_c del grupo (c) en una primera tabla S, donde la N-tupla (s_1, s_2, \dots, s_N) y el identificador del grupo I_c del grupo (c) asociado forman un elemento de dicha tabla S;
 - 20 ii) para cada elemento de la tabla S, identificar apariciones de los rellenos de ranura (s_1, s_2, \dots, s_N) en una pluralidad de artículos del grupo (c) y para cada aparición así identificada, almacenar los rellenos de ranura (s_1, s_2, \dots, s_N) junto con la frase en la cual ocurre en una segunda tabla C_0 ;
 - iii) de las frases almacenadas en la tabla C_0 , extraer patrones que abarquen los correspondientes N rellenos de ranura (s_1, s_2, \dots, s_N) dicho patrón extraído que expresa una relación semántica entre dichos N rellenos de ranura; y
 - 25 iv) almacenar dichos patrones extraídos junto con dicho patrón de entrada como una relación de implicación en dicha base de conocimiento.

Descripción de las realizaciones preferidas

El algoritmo descrito aquí es un algoritmo de aprendizaje automático iterativo de arranque que permite la intervención humana (en el paso 9). La meta del algoritmo es producir un conjunto de patrones, que parafrasean los patrones dados como entrada. El método explota grupos de documentos similares. Dado que la intervención humana es opcional el algoritmo puede trabajar tanto en el modo sin supervisión como en el modo supervisado.

Paso 1: El algoritmo acepta en sus entradas uno o más patrones p lineales o sintácticos (o un conjunto de P de tales patrones) con N ranuras (N>1) que expresan una relación R semántica específica. Por ejemplo, dos patrones sintácticos para la relación "criticar" pueden ser:

- 35 $PERSONA_1 \leftarrow \text{sujeto} - \text{criticó} - \text{objeto} \rightarrow PERSONA_2$
- $PERSONA_1 \leftarrow \text{sujeto} - \text{es_un_oponente} - \text{de} \rightarrow PERSONA_2$

Los patrones sintácticos son árboles sintácticos, que tienen dos o más ranuras. Cada ranura puede rellenarse con un tipo de entidad específico. En el ejemplo anterior, las ranuras son designadas como $PERSONA_1$ y $PERSONA_2$; pueden rellenarse solo por nombres del tipo persona. Las plantillas sintácticas se hacen corresponder contra árboles sintácticos, así no son dependientes del orden de palabras lineal y las palabras adicionales que puedan aparecer entre ellas. Por ejemplo, el primer patrón se corresponderá con frases como "John Adams, presidente de ABC Corp, criticó agudamente a su colega Adam Howard."

El algoritmo puede aceptar también patrones lineales como

$PERSONA_1$ criticó $PERSONA_2$

- 45 Diferente a los patrones sintácticos, los lineales dependen del orden de las palabras y las palabras por medio. Por ejemplo, el patrón lineal anterior hace corresponder solo frases como "John Adams criticó a Adam Howard", pero no hará corresponder "John Adams criticó agudamente a Adam Howard".

Paso 2: Otra entrada para el algoritmo de aprendizaje de patrones es un conjunto C de grupos c de artículos a. En cada grupo uno o más artículos deben estar presentes cuyo tema principal sea el mismo. El agrupamiento se puede hacer usando la semejanza de palabras entre artículos. Se describen diferentes algoritmos de agrupamiento en la literatura (véase por ejemplo "Frequent term-based text clustering" de Beil, Ester, y Xu).

- 50

Paso 3: a) Los artículos son procesados lingüísticamente con respecto al tipo de patrones a ser aprendidos. Si vamos a aprender árboles sintácticos, los artículos deberían ser analizados sintácticamente.

5 b) Las entidades que se corresponderán con el tipo de las ranuras del patrón deberían ser identificadas. Si tomamos como ejemplo los patrones en el Paso 1, entonces tenemos que identificar todos los nombres de persona. Se pueden usar herramientas para reconocimiento de Entidades de Nombres (NE) o diccionarios como WordNet (<http://wordnet.princeton.edu>) en este punto.

Paso 4: Los patrones p en el conjunto P se hacen corresponder contra los artículos en el conjunto C de grupos c . Un patrón se corresponde con un fragmento de texto o un árbol sintáctico, si todas sus partes se corresponden y todas las N ranuras se corresponden con entidades de texto del tipo correspondiente.

10 Cuando un patrón p de P se corresponde con un fragmento de texto en algunos artículos a que pertenecen al grupo c de C , el algoritmo almacena en una tabla S los N rellenos de ranura (s_1, s_2, \dots, s_N) que se corresponden con las ranuras del patrón y los identificadores I_c del grupo de c . Dado que los patrones p de P expresan la relación R , esta relación se mantendrá para (s_1, s_2, \dots, s_N) .

15 Paso 6: Para cada elemento de S , consistente en la N -tupla (s_1, s_2, \dots, s_N) de rellenos de ranuras y una referencia I_c a un grupo c de artículos a , buscamos todas las apariciones de los rellenos de ranuras en todos los artículos del grupo c . Consideramos solo las co-ocurrencias de (s_1, s_2, \dots, s_N) en una frase. Para cada tal co-ocurrencia creamos una fila en una nueva tabla C_0 donde almacenamos la N -tupla de rellenos de ranuras y la frase donde co-ocurre.

20 Note que los rellenos de ranuras fueron inicialmente extraídos de solo un artículo a de c y ahora buscamos su co-ocurrencia en todo el grupo c . Como hemos señalado en el punto 5, R mantiene los rellenos de ranuras en a . Además, dado que todos los artículos en un grupo se refieren al mismo tema, es razonable asumir que la relación R se mantiene para la N -tupla de relleno de ranura en todo el grupo c . Por ejemplo, si en un artículo del grupo c se enuncia que "John Adams criticó a Adam Howard", es razonable asumir que en la mayoría de los casos cuando John Adams y Adam Howard aparezcan juntos en algún artículo del grupo c , están relacionados a través de la relación "criticar".

25 Paso 7: A partir de las frases en la tabla C_0 se extraen patrones (lineales o sintácticos) que abarcan los correspondientes rellenos de ranuras N . Cada relleno de ranura en el patrón se sustituye por una designación de ranura. El tipo de cada ranura es igual al tipo del relleno de ranura. Diferentes enfoques para la extracción de patrones (tanto para patrones lineales como sintácticos) se describen en la literatura (véase por ejemplo "Scaling Web-based Acquisition of Entailment Relations" de Szpektor, Tanev, Dagan, y Coppola).

30 Paso 8: Cada patrón extraído en el paso 7 es pesado considerando el número de frases y el número de tuplas de relleno de ranura de C_0 que lo soportan. Una frase o una tupla de relleno de ranura soporta un patrón cuando el patrón aparece en la frase o abarca la tupla de relleno de ranura.

35 Paso 9: Los patrones cuyos pesos estén por encima de un cierto umbral (absoluto o relativo con respecto al peso más alto) son añadidos al conjunto P de patrones, si no existen allí. Si es relevante, un experto humano puede evaluar la oportunidad de los nuevos patrones añadidos.

Paso 10: Si el conjunto de patrones P no fue expandido en 9 o su tamaño está por encima de un cierto umbral, entonces para la iteración y se devuelve P como un resultado. Si no va al paso 4.

REIVINDICACIONES

1. Un método para formar relaciones de implicación; comprendiendo proporcionar un dispositivo informático y
 - a) proporcionar a dicho dispositivo informático al menos un patrón (p) de entrada con N ranuras de patrón ($N > 1$), dicho patrón (p) de entrada que expresa una relación semántica específica entre N entidades que rellenan las N ranuras de patrones del patrón (p) de entrada como rellenos de ranuras,
 - b) proporcionar a dicho dispositivo informático al menos un grupo (c) de artículos, dichos artículos de dicho grupo (c) relativos a un tema principal común;
 - c) procesar, por dicho dispositivo informático, dichos artículos con respecto al patrón (p) de entrada e identificar las identidades que se corresponden con el tipo semántico de las N ranuras de patrón;
 - d) si dicho al menos un patrón de entrada se corresponde con una porción de un artículo (a) de dicho al menos un grupo (c):
 - i) almacenar los N rellenos de ranura (s_1, s_2, \dots, s_N) que se corresponden con las ranuras del patrón (p), y un identificador I_c del grupo (c) en una primera tabla S, donde la N-tupla (s_1, s_2, \dots, s_N) y el identificador I_c de grupo del grupo (c) asociado forman un elemento de la tabla S;
 - ii) para cada elemento de la tabla S, identificar, mediante dicho dispositivo informático, apariciones de los rellenos de ranura (s_1, s_2, \dots, s_N) en una pluralidad de artículos del grupo (c) y para cada aparición de esta forma identificada, almacenar los rellenos de ranura (s_1, s_2, \dots, s_N) junto con la frase en la cual ocurren en una segunda tabla C_0 ;
 - iii) a partir de la frase almacenada en la tabla C_0 , extraer, mediante dicho dispositivo informático, patrones que abarquen los correspondientes N rellenos de ranura (s_1, s_2, \dots, s_N), dicho patrón extraído que expresa una relación semántica entre dichos N rellenos de ranura, donde cada relleno de ranura es sustituido por una variable de ranura; y
 - iv) formar, mediante dicho dispositivo informático, una relación de implicación entre dichos patrones extraídos y dicho patrón de entrada.
2. El método según la reivindicación 1, donde dicho patrón (p) de entrada es un patrón lineal.
3. El método según la reivindicación 1, donde dicho patrón (p) de entrada es un patrón sintáctico y donde dichos artículos son analizados sintácticamente antes del procesamiento de dichos artículos con respecto al patrón (p) de entrada en c).
4. El método según cualquiera de las reivindicaciones 1 a 3, donde un conjunto P de patrones (p) de entrada con N ranuras de patrón ($N > 1$) se proporciona, y donde el patrón extraído se añade al conjunto P de patrones.
5. El método según cualquiera de las reivindicaciones 1 a 4, donde un conjunto (C) de grupos (c) de artículos (a) se proporciona como entrada, y donde cada uno de dichos grupos (c) contiene uno o más artículos relativos a un tema principal común.
6. El método según cualquiera de las reivindicaciones 1 a 5, donde en ii) las apariciones de los rellenos de ranura (s_1, s_2, \dots, s_N) se identifican a lo largo de todos los artículos del grupo (c).
7. El método según cualquiera de las reivindicaciones 1 a 6, donde antes de iv), cada patrón extraído es pesado con respecto al número de frases y el número de N-tuplas de relleno de ranura que soportan el patrón extraído respectivo, y donde solo los patrones extraídos, para los cuales el peso exceda un umbral predeterminado, son además considerados en iv).
8. El método según cualquiera de las reivindicaciones 1 a 7, además comprende una evaluación de la oportunidad de la relación de implicación entre dicho patrón extraído y dicho patrón de entrada por intervención humana.
9. El método según cualquiera de las reivindicaciones 1 a 8, donde el paso d) además comprende el paso de:
 - iv) almacenar dichos patrones extraídos junto con dicho patrón de entrada como una relación de implicación en una base de conocimiento.
10. Un producto de programa informático almacenado en un medio usable por un ordenador, caracterizado por instrucciones operables para causar que un procesador programable lleve a cabo un método según cualquiera de las reivindicaciones 1 a 9.
11. Un programa informático que comprende código de programa para realizar todos los pasos de un método según cualquiera de las reivindicaciones 1 a 9 donde dicho programa se ejecuta en un ordenador.

12. El dispositivo informático que comprende

medios de memoria que tiene un código de programa almacenado en él para realizar todos los pasos de un método según cualquiera de las reivindicaciones 1 a 9, y

5 un entorno de ejecución para ejecutar dicho código de programa para llevar a cabo un método según cualquiera de las reivindicaciones 1 a 9.