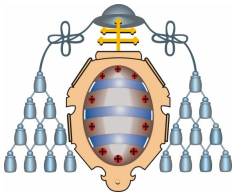
UNIVERSIDAD DE OVIEDO
Departamento de Informática



TESIS DOCTORAL

blindLight
Una nueva técnica para procesamiento de texto no estructurado mediante vectores de *n*-gramas de longitud variable con aplicación a diversas tareas de tratamiento de lenguaje natural

Presentada por

Daniel Gayo Avello

para la obtención del título de Doctor por la Universidad de Oviedo

Dirigida por el

Profesor Doctor D. Darío Álvarez Gutiérrez

Oviedo, Junio de 2005

RESUMEN

Es posible transformar, de manera automática, textos de cualquier idioma alfabético en vectores de n-gramas de longitud variable capaces de almacenar ciertos aspectos de la semántica subyacente al texto inicial. Estos vectores pueden transformar la información original, ser comparados e incluso combinados entre sí subrayando, como resultado, gran parte de la semántica presente en el texto de partida.

Se han utilizado frecuentemente n-gramas para llevar a cabo distintas tareas de tratamiento de lenguaje natural. La mayoría de estas técnicas tienen algunos puntos en común: (1) los documentos son mapeados sobre un espacio vectorial donde los n-gramas son utilizados como coordenadas y las frecuencias relativas de aparición de los mismos en el texto como pesos del vector, (2) muchas de estas técnicas producen un contexto para cada documento que juega un papel similar al de las listas de "palabras vacías" (stop-words) y (3) el coseno del ángulo formado por los vectores de los documentos se utiliza normalmente para determinar la similitud entre documentos o entre consultas y documentos.

blindLight es una nueva propuesta, desarrollada por este doctorando, relacionada con tales técnicas "clásicas" aunque introduce dos importantes diferencias: (1) no se utilizan las frecuencias relativas como pesos de los vectores sino las significatividades de los n-gramas y (2) se descarta el coseno del ángulo entre vectores de documentos en favor de una nueva métrica inspirada por las técnicas de alineación de secuencias aunque no tan costosa computacionalmente.

Esta nueva propuesta puede ser utilizada simultáneamente para categorizar u obtener grupos de documentos, recuperar información o extraer frases clave y resúmenes a partir de un único documento. Muchas de estas tareas son herramientas fundamentales para aliviar la "sobrecarga de información" y mejorar la experiencia de los usuarios.

ABSTRACT

It is possible to automatically transform texts written in any western language in variable-length n-gram vectors which preserve some of the semantics from the source texts. Such vectors can transform the primary information, be compared and even combined with each other highlighting, as a result, much of the semantics from the original document.

N-grams have been frequently used to perform different natural language processing tasks. Such methods show many features in common: (1) documents are represented using a vector space where n-grams are taken as coordinates and n-gram frequencies within documents as vector weights, (2) many of these techniques require a background which plays a role similar to that of lists of stop words and (3) the cosine similarity is normally used to compare documents to each other and documents to queries.

blindLight is a new approach, proposed by this researcher, related to such "classical" methods but with two major changes: (1) n-gram relative frequencies within documents are no more used as vector weights but their significances and (2) cosine distance is abandoned in favor of a new measure inspired by sequence alignment techniques although not so computationally expensive.

Such a new proposal can be used to perform automatic document clustering and categorization, information retrieval, in addition to keyphrase extraction and automatic summarization. Such tasks are essential tools to fight "information overload" and improve user experience.

AGRADECIMIENTOS

Mamá, va por ti.

Si he aprendido algo del largo proceso que me ha llevado hasta la conclusión de este trabajo han sido dos cosas: la investigación requiere rigurosidad y, sobre todo, humildad. He hecho cuanto estuvo en mi mano para conseguir lo primero y debo decir que espero de corazón que mi futuro trabajo sea mejor que esta limitada contribución.

Por otro lado, es sabido que un trabajo de esta índole nunca es posible en solitario (como atestiguan dieciocho páginas de referencias) y muchas personas han contribuido a que yo pudiera llevarlo a cabo. Mencionar a todos y cada uno de los que, de un modo u otro, han participado en mi formación académica y profesional resultaría demasiado extenso y podría olvidarme de alguien, así que lo haré de manera breve y sin entrar en detalles, cada uno de ellos sabe a qué me refiero.

En primer lugar me siento agradecido a los hombres y mujeres que forman el Departamento de Informática de la Universidad de Oviedo, desde la dirección al PAS, pasando por compañeros y alumnos. Todos ellos hacen que éste sea un lugar en el que, a pesar de las adversidades, siga siendo una satisfacción y un orgullo trabajar. Un recuerdo especial para Brugos, Cobas y María Jesús que tanto me ha ayudado con el papeleo.

Naturalmente, en un departamento tan grande es imposible tener la misma familiaridad con todo el mundo y si hay un grupo donde me he sentido acogido (y protegido) es Oviedo3, un abrazo para Cueva y Benjamín y un beso para Marián y Almudena, sin ellas Introducción a la Programación me habría sobrepasado.

No puedo dejar de recordar a los compañeros con los que compartí durante cuatro cursos "El Palomar". La incomodidad, el calor y la falta de espacio se olvidaban gracias a vosotros. De entre ellos debo citar a los dos caballeros que durante ese tiempo formaron conmigo un 3 en raya humano: Guti y Luis. Cada conversación que tuvimos fue un placer y mucho de lo que hablamos me ayudó a terminar este trabajo o me confortó en momentos difíciles. A Luis debo agradecerle tantos correos inspiradores y aquella traducción al sueco. A Guti he de agradecerle muchas cosas pero sólo mencionaré una: su libro acerca de cómo escribir una tesis, debes terminarlo.

Mención aparte merece Darío, director de esta tesis, que me ha llevado casi de la mano por los difíciles caminos de la investigación y que mucho antes demostró una confianza en mí por la que siempre le estaré agradecido.

Pero por encima de todo tengo una deuda infinita con dos mujeres que ahora mismo se sienten muy aliviadas: Tensi y mi madre. Ambas ocupan ex aequo mi corazón y no creo que exagere si digo que ellas han sufrido con esta tesis más que yo y que, en consecuencia, es más un logro suyo que mío. Por todo vuestro apoyo, amor y, sobre todo, paciencia, ¡gracias! Aunque hubiese sido capaz de terminar sin esas tres cosas no habría merecido la pena.

TABLA DE CONTENIDOS