CAPÍTULO 8

CONCLUSIONES Y TRABAJO FUTURO

El tema central de este trabajo es la "sobrecarga de información" y la forma de afrontarla. El autor está interesado en una forma particular de información: texto libre escrito en cualquier lenguaje natural; y en un entorno específico en que se produce dicha sobrecarga: los medios *online*. Esta sobrecarga de información es un problema antiguo que se ha visto acrecentado por el progreso tecnológico: por un lado, el abaratamiento de los soportes físicos permite almacenar textos sin tener que plantearse la eliminación de ningún documento no importa lo obsoleto o inútil que sea (véase Fig. 123) y, por otro, la disponibilidad de estándares de comunicación, transporte, formateo de documentos, etc. que han permitido que prácticamente cualquier usuario pueda convertirse en fuente de información.

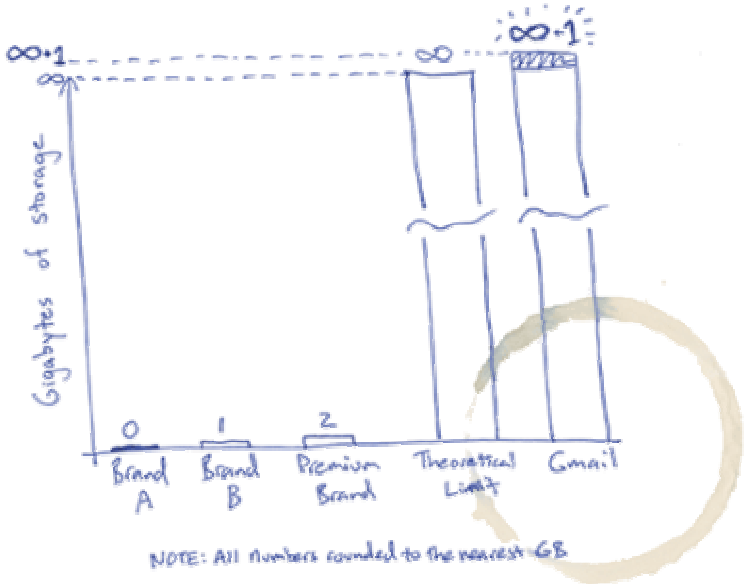

Fig. 123 Una "broma seria" de *Google* que ofrece a los usuarios de la solución de correo electrónico *GMail* una cuota creciente y presuntamente ilimitada a partir de 2GB.

El autor realizó una propuesta para afrontar dicho problema: la Web Cooperativa que se sustenta sobre tres puntos:

1. La utilización de conceptos, generados automáticamente, como alternativa intermedia entre las ontologías y las palabras clave.
2. La clasificación de documentos en una taxonomía a partir de tales conceptos.
3. La cooperación entre usuarios, en realidad, entre agentes que actúan en representación de los usuarios y que no requieren su participación explícita.

El primer punto venía motivado por los inconvenientes de la utilización de palabras clave para la formulación de las "necesidades de información" por parte de los usuarios y la dificultad para desarrollar ontologías que diesen soporte a cualquier consulta concebible. El autor planteó los conceptos como entidades más abstractas y, por tanto, con mayor carga semántica que las palabras clave pero que, al mismo tiempo, pudiesen ser obtenidos de manera automática. Una posible forma de construir tales conceptos sería mediante "agrupaciones débiles" de términos relacionados y una tecnología que podría llevar a cabo esta tarea es la semántica latente aunque también podría utilizarse la técnica presentada como parte central de este trabajo.

Frente a la Web Semántica que requiere la utilización de un "marcado ontológico" la Web Cooperativa propone la utilización de una semántica más sencilla[1]: simples categorías obtenidas de manera automática a las que los distintos documentos podrían asociarse sin necesidad de emplear ningún tipo de etiquetas. Para ello el autor sugiere utilizar tan sólo el texto plano de los documentos, argumentando que éstos pueden considerarse individuos de una población mayor y que, del mismo modo en que es posible clasificar y categorizar de manera automática a los seres vivos mediante su código genético, es posible adaptar algoritmos empleados en biología computacional al campo de la clasificación de documentos.

Por último, con la Web Cooperativa se pretende aprovechar el conocimiento experimental que obtienen los usuarios al explorar la Web actual y que se desaprovecha en gran medida. Para ello se pretende utilizar técnicas que permitan obtener de los usuarios, de forma no intrusiva y transparente, información sobre la relevancia de los distintos documentos. Así, cada usuario de la Web Cooperativa dispondría de un agente con dos objetivos: aprender de su "maestros" y recuperar información para él.

Por tanto, el objetivo último de la Web Cooperativa sería, por un lado, dotar de una cierta semántica a la Web mediante técnicas automáticas y explotar la experiencia diaria de los usuarios en provecho de los mismos: facilitándoles la ejecución de consultas adaptadas a sus necesidades y ofreciéndoles, sin precisar de consulta alguna, información relevante.

Así pues, la Web Cooperativa involucra muy diversos aspectos: tratamiento de lenguaje natural, evaluación implícita de documentos, agentes *software*, interacción persona-ordenador, usabilidad o privacidad. Por otro lado, sus objetivos son muy ambiciosos y exceden con creces no sólo los fines de un trabajo como esta disertación sino, sobre todo, las capacidades de un único investigador. Por esa razón se plantearon una serie de cuestiones previas para determinar que subconjunto de las mismas constituían, a un

[1] Desde hace algún tiempo existe una *meme* que también defiende el uso de etiquetas más sencillas pero aún dotadas de semántica: "folksonomía". Este término, del inglés *folksonomy* = *folk* + *taxonomy*, hace referencia a la categorización colaborativa de documentos mediante etiquetas (palabras clave) escogidas libremente por los usuarios. Aún nos encontramos en una fase muy incipiente como para predecir el desarrollo de estos métodos pero será necesario prestarles atención.

tiempo, un problema interesante y practicable. Las preguntas finalmente escogidas fueron las siguientes:

- ¿Es posible clasificar textos libres empleando métodos tomados de la biología computacional?
- ¿Es posible obtener un "pseudo-ADN" a partir de texto escrito en un lenguaje natural?
- Si existiera ese pseudo-ADN, ¿sería posible combinarlo, mutarlo o construir "algo" a partir del mismo?
- ¿Se debe suponer que idiomas distintos constituyen "bioquímicas" diferentes?
- ¿Cómo se daría el salto desde ese pseudo-ADN a los conceptos?
- ¿En qué forma podría un agente realizar búsquedas eficientes sobre una taxonomía de documentos construida a partir de ese pseudo-ADN?

Así, para la realización de este trabajo se prescindió de lo que serían las "capas superiores" de la Web Cooperativa y se delimitó mejor el problema a resolver:

> *La cantidad de texto no estructurado disponible en la Web seguirá aumentando y, a pesar de sus inconvenientes, el método preferido por la mayor parte de usuarios para recuperar información continuarán siendo las consultas formuladas en lenguajes naturales. En ambos casos (publicación y consulta) será inevitable un uso generalmente ambiguo de los distintos idiomas y la presencia de errores tipográficos, ortográficos o gramaticales.*

De este modo se encuadró el problema dentro del campo de procesamiento de lenguaje natural por medios estadísticos y se formuló la siguiente tesis:

> *Se puede obtener para los distintos n-gramas, g_i, de un texto escrito en cualquier idioma una medida de su significatividad, s_i, distinta de la frecuencia relativa de aparición de los mismos en el texto, f_i, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, d_i, un único vector, v_i, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir. Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual, facilitando la clasificación y categorización de documentos, así como la recuperación de información. Asimismo, cada vector individual es capaz de transformar el texto original a partir del cual fue obtenido dando lugar a secuencias de palabras clave y resúmenes automáticos.*
>
> Que se resumía de este modo:
>
> *Una única técnica sencilla, basada en el uso de vectores de n-gramas de longitud variable, independiente del idioma y aplicable a diversas tareas de tratamiento de lenguaje natural con resultados similares a los de otros métodos 'ad hoc' es viable.*

En resumen, aunque el trasfondo de este trabajo es la sobrecarga de información el problema de que realmente se ocupa es el del **procesamiento de texto natural por medios estadísticos y en condiciones "extremas"**: multilingüismo, gran número de documentos, textos ambiguos, no estructurados y con ruido (errores tipográficos, ortográficos o gramaticales).

Este problema, especialmente notable en los entornos *online*, puede desglosarse en una serie de tareas: asignación de documentos a categorías conocidas (**categorización**),

agrupación de documentos con características similares (**clasificación o *clustering***), **recuperación de información** y destilación de información (p.ej. **resumen automático**).

No es necesario decir que existen diversas técnicas capaces de afrontar una o más de las tareas anteriores. Sin embargo el autor planteó que una única técnica bastaba para resolver todas las tareas de manera adecuada verificando, además, las siguientes características: independencia del idioma, utilización de métodos puramente estadísticos y alta tolerancia al ruido.

Dicha técnica, propuesta por el autor y denominada *blindLight*, es una técnica parcialmente bioinspirada puesto que parte del concepto de "ADN documental" que estaría formado por una secuencia de genes, pares constituidos por un *n*-grama de caracteres y la significatividad de dicho *n*-grama dentro del texto del documento.

La principal diferencia entre esta propuesta y otras también basadas en la idea de un "genoma documental" radica en que no se pretende emplear éste únicamente para categorizarlo o clasificarlo sino que, al igual que el ADN de los seres vivos, este "ADN documental" puede combinarse entre sí además de "activarse" produciendo un resultado diferente del texto original.

Las razones para emplear *n*-gramas de caracteres que pueden "saltar" entre palabras son varias: (1) facilitan un trato "igualitario" a todos los lenguajes aun cuando no utilicen separadores de palabras (p.ej. el chino o el japonés), (2) garantizan una tolerancia elevada al ruido, (3) permiten obviar el uso de ciertos algoritmos (p.ej. los de *stemming*) y ofrecen un rendimiento adecuado incluso en idiomas muy complejos (p.ej. el finés).

Por lo que respecta al cálculo de la significatividad de cada *n*-grama en el documento se pueden emplear toda una serie de estadísticos (p.ej. información mutua, Dice, χ^2, probabilidad condicional simétrica, etc.) que no requieren la utilización de un contexto en que situar al documento (algo necesario si se utilizase, por ejemplo, *tf*idf*).

Por otro lado, en tanto que cadenas, en particular de inspiración biológica, parecía claro que sería posible aplicar algoritmos tomados de la biología computacional y llevar a cabo la clasificación automática de documentos representados de este modo. Sin embargo, era precisa una solución más general que permitiese determinar la similitud entre dos cadenas de este "ADN documental".

Antes de definir esa medida de asociación se describió un proceso de "hibridación", o mejor, intersección entre cadenas de pseudo-ADN de tal modo que a partir de dos de tales cadenas se obtuviese una tercera. En términos biológicos se puede afirmar que cuanto mayor sea la longitud de la cadena híbrida más elevado resulta el grado de parentesco entre las dos cadenas originales. De modo análogo, *blindLight* define una operación de intersección de cadenas de "ADN documental" y, en consecuencia, establece dos medidas asimétricas denominadas Π (Pi) y P (Ro) que vinculan la significatividad total de la cadena intersección o híbrida con la de cada uno de los dos progenitores.

Dependiendo de la longitud de los documentos a comparar las significatividades totales de ambos y de la cadena intersección pueden ser muy distintas y, en consecuencia, los valores de Π y P muy diferentes. No obstante, el hecho de que sean independientes permite su combinación lineal de diversas formas y su adaptación a las diversas necesidades de cada tarea.

Uno de los aspectos que pueden resultar más controvertidos de la tesis del autor es la afirmación de que los vectores de *n*-gramas empleados para representar este pseudo-ADN

son capaces de almacenar aspectos semánticos subyacentes al texto original, en particular, si se tiene en cuenta la pretensión de que la técnica es válida para cualquier tipo de lenguaje natural. No obstante, una serie de experimentos relativos a la clasificación de traducciones literales de un conjunto de documentos en español, inglés, francés, finés, holandés, hebreo y japonés así como el resumen de textos en inglés, alemán, francés, danés y húngaro concluyeron que ***blindLight***, aun cuando no sea total y absolutamente independiente respecto al idioma, **muestra un comportamiento extremadamente consistente entre lenguajes muy diferentes** por lo que se puede afirmar que, efectivamente, este "ADN documental" sí almacena ciertos aspectos semánticos del texto. Una vez garantizado este aspecto de la técnica se describió su aplicación a cada una de las tareas anteriormente descritas: clasificación, categorización, recuperación de información y destilación de información (en particular extracción de resúmenes y palabras clave).

Para la primera tarea, la **clasificación automática de documentos**, se presentaron dos algoritmos basados en *blindLight* (uno incremental y otro no incremental) y se comparó la nueva técnica propuesta con otros métodos, resultando que ***blindLight* ofrece un rendimiento similar al de ciertas técnicas** (p.ej. mapas auto-organizativos) **y mejor que el de otras como los métodos particionales y jerárquicos**. También se llevó a cabo un experimento relativo a la clasificación genética de lenguajes naturales empleando textos de 14 idiomas europeos y transcripciones fonéticas de 9 idiomas. Los resultados de ambas clasificaciones no sólo fueron coherentes entre sí sino también con las teorías lingüísticas vigentes. A raíz de estas experiencias se concluyó que, en efecto, era posible emplear *blindLight* como técnica de clasificación automática con resultados análogos a los de métodos específicos.

Posteriormente se describió la aplicación de la técnica propuesta por el autor a la **categorización de documentos** y se llevaron a cabo una serie de experimentos relativos a identificación de idiomas, autoría de documentos, filtrado de correo no deseado así como una prueba con las colecciones *Reuters 21578* y *OHSUMED*. A la luz de tales experimentos se puede concluir:

1. La utilización de *blindLight* como sistema para la identificación de idiomas proporciona unos resultados muy similares (y bajo ciertas condiciones de longitud del texto y ruido superiores) a técnicas reconocidas aun empleando apenas 10KB de información para cada idioma.
2. La aplicación de *blindLight* como filtro de *spam* requiere mejoras pero como simple experimento de categorización parece sugerir un rendimiento similar al de los clasificadores Bayesianos y *MBL* (*Memory Based Learning*).
3. Las pruebas estandarizadas indican que *blindLight* ofrece resultados análogos a clasificadores Bayesianos, Rocchio y árboles de decisión, apreciablemente inferiores a *k*-vecinos y sustancialmente inferiores a las *SVM* (*Support Vector Machines*).

En resumen, ***blindLight* no alcanza los resultados de las *SVM* pero es similar a otras técnicas empleadas con frecuencia y aceptadas como adecuadas** (p.ej. clasificadores Bayesianos).

Para la evaluación de *blindLight* como técnica de **recuperación de información** se experimentó con las colecciones *CACM* y *CISI* y se tomó parte en la edición de 2004 del *CLEF*. Es necesario decir que los resultados obtenidos en ambos casos fueron claramente **inferiores a los de las técnicas tradicionales** por lo que la afirmación respecto a la viabilidad de esta técnica para su utilización en *IR* queda, por el momento, en suspenso. No obstante, es necesario señalar una serie de aspectos alentadores y que sugieren que, en el

futuro, tal vez sea posible situar a *blindLight* al mismo nivel que otras técnicas consolidadas. En primer lugar, aunque las técnicas tradicionales de recuperación de información superan a la del autor, otras nuevas y consideradas "prometedoras" (como el indexado por semántica latente) ofrecen resultados muy similares; por otro lado, las medidas de similitud entre consultas y documentos pueden mejorarse (tal vez empleando programación genética) y uno de los elementos empleados en *CLEF'04* (el sistema de pseudo-traducción) aún estaba en una fase preliminar.

Por lo que respecta a la **extracción de resúmenes** y palabras clave se describió el modo en que se puede utilizar el "ADN documental" para segmentar el texto plano del documento en fragmentos de máxima significatividad empleando un procedimiento inspirado en la síntesis de las proteínas. Este método permite obtener una información muy valiosa para determinar las sentencias más relevantes del texto y construir así un resumen extractivo. Para evaluar este enfoque se emplearon los datos de la edición 2004 de *DUC* (*Document Understanding Conferences*) obteniendo unos resultados muy alentadores: al extraer resúmenes cortos (máximo 665 caracteres) a partir de un conjunto de documentos, ***blindLight* resultó ser superior a muchas de las técnicas más avanzadas disponibles**, aunque aún está lejos de alcanzar a los mejores sistemas existentes en la actualidad.

Por lo que respecta al desarrollo futuro de este trabajo existen varias líneas interesantes:

1. Adaptar el sistema de extracción de resúmenes a entornos multidocumento.
2. Continuar el desarrollo del sistema de pseudo-traducción.
3. Analizar la posible integración de los dos sistemas anteriores.
4. Emplear programación genética para la obtención de nuevas medidas de similitud entre documentos y consultas en el sistema *IR*.
5. Estudiar la posible integración de medidas basadas en la complejidad de Kolmogorov.
6. Estudiar la utilización de fragmentos de significatividad máxima como términos de indexado en el sistemas *IR*.

En conclusión, el autor ha presentado una **técnica novedosa** para el **procesamiento de lenguaje natural** por medios puramente estadísticos. Dicha técnica es aplicable a **múltiples idiomas** ofreciendo resultados consistentes en todos ellos, muestra una adecuada **tolerancia al ruido** y resulta **apta para tareas de clasificación, categorización y extracción de resúmenes**. Además, parece **potencialmente útil** para la **recuperación de información** en entornos **multilingües** aunque en este campo aún no se ha progresado lo suficiente.