ÍNDICES MULTIMEDIA: EL FUTURO DE LA WEB

Edgar Chávez[1], Antonio Camarena[2] y Eric Sadit Téllez[3]

Universidad Michoacana de San Nicolás de Hidalgo

(1) elchavez@umich.mx (2) camarena@umich.mx (3) sadit@lsc.fie.umich.mx

MESA: Las Ciencias Exactas y Materiales

Resumen

La red global de datos (World Wide Web o WWW por sus siglas en inglés) ha sido el cambio más disruptivo en la historia de la humanidad. En menos de veinte años las computadoras se han vuelto pervasivas y transversales; no existe prácticamente ninguna actividad humana que no utilice de una manera u otra las computadoras. Si una computadora es omnipresente en la vida cotidiana, el internet se ha convertido en la biblioteca global y en solo 10 años las empresas dedicadas a proveer servicios de búsqueda en repositorios de información conectados a internet se han convertido en las mas rentables del mundo, notablemente *google* es omnipresente en la vida de miles de millones de personas alrededor del mundo; superando en capital a empresas que proporcionan bienes tangibles.

A pesar de que la revolución anterior ya es sorprendente y disruptiva, en el mundo una nueva mini-revolución está gestándose; también alrededor de las tecnologías de búsqueda; esta vez, en la búsqueda de contenido multimedia. El esfuerzo científico y tecnológico en ciencia y tecnología de la información se ha orientado a dotar a los buscadores de capacidades para buscar objetos multimedia por contenido.

Tanto la búsqueda en texto (que describe las capacidades actuales de los buscadores en internet) como la búsqueda multimedia (que describe las capacidades futuras de los buscadores) son el resultado de la aplicación de modelos matemáticos discretos y continuos. Los algoritmos son el corazón de esta revolución. A grosso modo, la limitación actual de los algoritmos de búsqueda multimedia consiste en su falta de escalabilidad, a diferencia de sus contrapartes de texto que pueden buscar en miles de millones de páginas.

En este artículo presentamos un conjunto de algoritmos que permiten hacer búsqueda por contenido en repositorios multimedia en tiempo constante; es decir, de manera super-escalable, de la misma manera que búsqueda en texto.

Introducción

Desde antes de la existencia de las computadoras se ha soñado con organizar el cúmulo de datos, información y conocimientos que la humanidad posee. Quizá la primera propuesta precisa fue establecida en 1934 por Paul Otlet (considerado el padre de las ciencias de la información, creador del sistema digital de clasificación bibliotecaria, de las fichas bibliográficas, los cardex, etc.) quien propuso la creación de un repositorio de información accesible por todo el mundo a través de un mecanismo que era una extrapolación del telégrafo y con una representación sucinta de los documentos en diversas bibliotecas en los cardex de cada una. Este pionero pudo visualizar las ciudades del conocimiento, la web, las redes sociales y la aldea mundial basada en tecnologías analógicas. Cincuenta y cinco años después la propuesta de Tim Berners-Lee, que dio lugar a la web desde el CERN, y la posterior proliferación de las máquinas de búsqueda han permitido cumplir parcialmente la expectativa de tener a la mano la información producida por toda la humanidad.

La noción de Sociedad de la Información tiene sus orígenes en la economía, al tratar de encontrar el motor del desarrollo en cosas menos tangibles que la producción de bienes, con los trabajos pioneros de Fritz Machlup. Ahora es evidente que estamos instalados plenamente en una sociedad en donde el desarrollo depende de la capacidad de acumular, procesar y asimilar información; pero aparentemente solo estamos conscientes de una cara de la moneda: nuestro papel como consumidores de información. Desde la educación formal (escolar), hasta la educación informal que da la interacción social en su conjunto lo que hacemos es consumir información. Una cosa de la que no estamos completamente conscientes es que nosotros producimos información, o de manera más precisa, producimos datos que posteriormente pueden traducirse en información. Con el abaratamiento de los dispositivos de almacenamiento y los avances en los dispositivos de captura podemos almacenar audio y video digitales, tener información multimedia, mapas mundiales al alcance de la mano, a unas cuantas teclas de distancia, de manera instantánea, desde nuestro escritorio compartidas con el resto de la sociedad.

Todos los avances que hemos presenciado son solo una parte de lo que se puede hacer con el cúmulo de datos que existen en la telaraña mundial de redes de computadoras. Hay una parte sutil; pero sustancial, que aguarda por una solución. La humanidad está en el vértice de un cambio tan dramático como el que ha provocado la web. Hemos llegado a un punto en el que podemos tener memorizada toda la actividad humana. El rango de las cosas que se almacenan va desde un paseo por el parque donde se toman fotos y video hasta la secuenciación del ADN, datos astrofísicos, colisiones de partículas elementales, tomografías, placas radiológicas, seguido de un largo etcétera.

El hecho de poder almacenar los datos no implica que tengamos información de ellos y de la misma manera, tener la información no implica tener el conocimiento. Finalmente, el conocimiento nos permite tomar decisiones.

Tomemos como ejemplo el conocimiento que tenemos acerca de las enfermedades infecciosas. Sabemos que estas enfermedades son causadas por bacterias o virus y eso permite tomar la decisión (automática) de esterilizar el instrumental en una operación, fijar nuestra atención en el desarrollo de vacunas o antibióticos. Los datos que permitieron tener la información necesaria para el conocimiento anterior fueron observaciones de personas enfermas, placas del microscopio, epidemias, etc.

En esta llamada era del *conocimiento*, los países que sobresalen son los capaces de generar conocimiento. Una forma de generar conocimiento es mediante el desarrollo y aplicación de técnicas modernas de análisis de información.

México tiene la oportunidad y la obligación de posicionarse como país generador y exportador de una de las tecnologías que van a moldear el futuro de la sociedad. Su desarrollo, sin embargo, requiere de una inyección de recursos adecuadamente enfocados, controlados y rigurosamente evaluados que permitan crear las condiciones necesarias en el país para impulsar el área, coordinando diferentes grupos, tanto de TI como de otras disciplinas, y formando recursos humanos que sustenten y le den continuidad al área.

El desarrollo de esta área permitirá realizar avances importantes en diferentes aspectos de las ciencias computacionales, contribuirá al desarrollo de diversas disciplinas que requieran utilizar datos, ayudará a mejorar la competitividad del país, mejorará aspectos en salud, permitirá a instancias gubernamentales tomar mejores decisiones, ayudará a integrar a una comunidad computacional interesada en estas áreas y fomentará la colaboración con instancias de gobierno, con la industria y con otras comunidades científicas.

Búsqueda de proximidad

Para establecer un vocabulario común pensemos en la definición formal del problema. Un espacio métrico es un par (X,d), donde X es un conjunto (los objetos válidos, e.g. las imágenes o las canciones), y d es una función de distancia, d : $X \times X \rightarrow R^{+}$ La función de distancia d(·,·) modela la semejanza entre objetos. X puede ser en principio infinito, un subconjunto finito de X será la base de datos, nuestro objeto de estudio. La mayoría de los índices requiere que d(·,·) sea una métrica, es decir que cumpla con las propiedades d(x,y) > 0 si x /= y, d(x,y) = d(y,x), d(x,x) = 0 y la desigualdad del triángulo: d(x, z) ≤ d(x, y) + d(y, z). Denotemos a la base de datos como S □ X un conjunto finito.
Una consulta es entonces de dos tipos básicos. Una *consulta de rango*, denotada por (q,r)d = {u □ S | d(q,u) ≤ r} para q □ X y r real. La otra consulta interesante es obtener los k elementos mas cercanos a q, denotado por KNN(q)d = {s □ S : □u □ Sd(q, s) ≤ d(q, u)} and |KNN(q)d| = k.
Los índices métricos tradicionales hacen uso de la desigualdad del triángulo para evitar cálculos de distancias.

Se puede probar que $|d(s,p)-d(q,p)| \leq d(s,q)$, mas aún; para un conjunto p = {pi} se cumple D(s, q) = Maxi|d(s, pi)−d(q, pi)|, de tal modo que (q, r)d □ (q, r)D. Esto implica que es posible resolver la consulta (q, r)d sin tener que verificar a toda la base de datos; únicamente hay que verificar los elementos de (q, r)D en cuya resolución solo se emplean las distancias de la consulta q a los elementos pi; mientras que las distancias d(s, pi) pueden ser precalculados y no se requieren al tiempo de consulta.

El método anterior funciona siempre que el espacio métrico tenga una dimensión intrínseca baja; de otro modo sufre la llamada *maldición de la dimensión*. De manera breve lo anterior significa que |(q,r)D| □ |S| tienen tamaños comparables, o de otro modo, que no acota mucho la consulta el utilizar el conjunto de referencia p.

En 2005 propusimos un método alternativo de solución del problema al guardar las *permutaciones* del conjunto de referencia en lugar de guardar directamente las distancias. La proximidad de los objetos en la base de datos se puede predecir midiendo la distancia entre permutaciones. Este método no sufre de la maldición de la dimensión aunque es un método aproximado; esto implica que el conjunto de respuesta en el índice no contiene a (q, r)d pero tiene una intersección grande con el.

La simplificación hecha por Google, Mufin o Mipai consiste en evaluar la distancia entre permutaciones mediante un índice invertido (Mufin) o un árbol de sufijos (Mipai) o Locality Preserving Hashing (Google). Este último es un método aplicable a vectores de dimensión muy alta; las permutaciones son un método para obtener vectores de objetos. Las tres simplificaciones implican pérdida en el recall. Nuestra propuesta en CIARP 2009 consiste en tener una representación sucinta de las permutaciones, como vectores binarios y encontramos que la distancia de Hamming es un excelente predictor para la distancia entre permutaciones. Al momento de escribir estas líneas estamos trabajando en un par de alternativas de representación de los objetos que prometen una eficiencia más alta que la de representación de permutaciones sin sacrificar el recall.

Caracterización sensorial de las señales de audio

Decimos que una señal de audio está caracterizada sensorialmente por una función de distancia *d* si cuando aplicamos *d* a pares de señales de audio que son sensorialmente iguales obtenemos valores pequeños de *d* y cuando la aplicamos a pares de señales sensorialmente distintas obtenemos valores grandes de *d*. Para fijar ideas podemos asumir que las señales de audio están alineadas, es decir; que al compararlas lo podemos hacer como vectores, componente a componente. Por simplicidad también podemos asumir que compararemos solo objetos completos. Estas dos condiciones se pueden relajar después para dar lugar a suposiciones más generales. De entre todas las posibles funciones que podrían caracterizar a las señales de audio, nos interesan aquellas que permitan crear índices eficientes. No cualquier representación tiene asociada una distancia natural que sea además indexable. En este sentido hemos tenido avances importantes en el desarrollo de representaciones indexables de señales de audio.

En 2008 el Dr. Antonio Camarena obtuvo el grado de doctor, desarrollando una representación de la señal de audio basada en la medición de la cantidad de información instantánea (en una ventana de tamaño fijo) contenida en la señal. Esta representación resulta ser muy estable, invariante a diversas deformaciones como diferencias en volumen, filtros pasa bajas, ruido blanco o coloreado, ecualización y compresión con pérdida. La técnica consiste en medir la entropía del histograma de la señal en una ventana de tiempo. Si el histograma en el intervalo [t1,t2] se denota como {pi}, normalizado para que se obtenga una distribución de probabilidad, la entropía en ese intervalo seria H([t1, t2]) = − □i pi log pi. Esta medida se toma en intervalos con traslape cubriendo toda la señal. La manera mas estable de obtener una representación que agrupe señales perceptualmente iguales consiste en tomar la entropía por bandas en la escala logarítmica de Bark. Nótese que en este contexto pi es la probabilidad de algún valor de la señal.

Hemos aplicado la representación por bandas de la entropía para la obtención de meta-datos (ej. nombre, grupo, genero, etc.) en consultas por contenido en repositorios de audio [4], para la identificación de comerciales en emisiones radiofónicas [3,1] y para el reconocimiento de *performances* o interpretaciones de piezas musicales por artistas distintos [2].

Los repositorios Mipai, Mufin o Images de Google permiten consultas que son holísticas; en el sentido de que se comparan imágenes completas. La función de distancia que se utiliza mide la respuesta cromática y es conocida como *características mpeg 7*. La estructura de las imágenes no se considera; quiere decir que dos tomas de la misma imagen en donde una esta en blanco y negro y otra en color no resultarán semejantes.

La representación de la señal de audio se convierte en una serie de tiempo, o en un conjunto de series de tiempo cuando se obtiene en múltiples bandas. Una consulta es un segmento de la serie de tiempo (una grabación de unos pocos segundos) y el objetivo es encontrar a que serie de tiempo corresponde el mejor empate con la consulta. Aquí es donde el uso de los índices resulta relevante. Una técnica que hemos utilizado con éxito para este problema, consiste en crear un índice con todos los posibles segmentos de la señal (sin traslape) y compararlos con todos los corrimientos de la señal. Con este método podemos consultar en un repositorio de 40,000 canciones con consultas de unos 5 a 10 segundos de longitud, obteniendo respuesta en menos de un segundo y con recall perfecto.

El grupo de la republica Checa ha trabajado en diferentes aplicaciones en la dirección de mufin *http://mufin.fi.muni.cz/tiki-index.php,* donde hay muchos demos de aplicación en diferentes dominios. Otro ejemplo de aplicación es *http://mipai.esuli.it/* donde se explora la búsqueda de imágenes por contenido. Quizá el más conocido es *http://images.google.com* donde se puede buscar por contenido muchas imágenes de la web.

Nuestro propio trabajo se compila en http://www.natix.org con demos de texto, imágenes y audio.

Referencias

[1] A. Camarena-Ibarrola, E. Cháv́ez. (2006). A robust entropy-based audio-fingerprint. *International Conference on Multimedia and Expo (ICME)*. :1729-1732.

[2] A. Camarena-Ibarrola, E. Chavez, (2006). On musical performances identification, entropy and string matching. *Fifth Mexican International Conference on Artifi- cial Intelligence 2006 (MICAI2006)*. **4293**:952-962.

[3] A. Camarena-Ibarrola, E. Chávez, E. Sadit. (2009). Robust Radio Broadcast Monitoring Using a Multi-Band Spectral Entropy Signature. Congreso Iberoamericano de Reconocimiento de Patrones, Springer LNCS. 5856.

[4] E. Sadit, E.Chavez, A Camarena-Ibarrola. (2009). A Brief Index for Proximity Searching. *Congreso Iberoamericano de Reconocimiento de Patrones, Springer LNCS.* **5856**.