



I dati della British Library sul web

IBM sta collaborando con la British Library ad un progetto per la conservazione e l'analisi di terabytes di informazioni sul Web prima che vadano persi per sempre.

Il nuovo progetto di analytics software, denominato IBM BigSheets, aiuta ad estrarre, annotare e analizzare

visivamente grandi quantità di informazioni Web utilizzando un browser. Il prototipo della nuova tecnologia IBM sta aiutando la British Library ad archiviare e preservare quantità massicce di pagine Web e, quindi, aprire la porta virtuale dei suoi archivi alle generazioni future.

La nuova tecnologia di IBM aiuta la British Library a velocizzare il processo di archiviazione, prima che il dato presente su Web venga perso per sempre. Il Web cambia velocemente, con nuove pagine che vengono create ogni giorno: un'esplosione di informazioni, destinate a sparire quasi altrettanto velocemente. Recenti ricerche stimano l'aspettativa di vita di un sito Web tra i 44 ed i 75 giorni. Ogni sei mesi, il 10% delle pagine Web dei domini inglesi viene perso.

"IBM BigSheets fa per le grandi quantità di informazioni quello che un foglio di calcolo ha fatto per un personal computer," ha detto Rod Smith, vice president, Emerging Internet Technologies, IBM. "Nello spazio di pochi minuti, ricercatori, accademici e studenti potranno eseguire ricerche su archivi di pagine Web di grandi dimensioni, analizzare e visualizzare senza sforzo i risultati della ricerca."

Preservare le informazioni per le generazioni future

Ogni anno più di sei milioni di ricerche vengono generate a partire dal catalogo online della British Library, e più di 400.000 persone visitano le sale-lettura della British Library, alla ricerca di informazioni. La British Library riceve una copia di ogni pubblicazione fisica prodotta nel Regno Unito ed in Irlanda, per un totale di 150 milioni tra mappe, manoscritti, spartiti musicali, giornali e riviste che deve archiviare. Andando oltre il semplice aspetto fisico, la British Library ha avviato l'archiviazione di pagine Web scelte dai domini UK a partire dal 2004. Con BigSheets, gli utenti della biblioteca, in futuro, avranno la possibilità di accedere ad un vasto archivio storico di siti Web e di fare ricerche e analisi, visualizzandone i risultati, in modo semplice.

"Stimiamo che lo spazio Web del Regno Unito conterrà, entro il 2011, oltre 11 milioni di siti Web. Per affrontare l'enorme sfida di catturare questi contenuti, abbiamo bisogno di un sistema capace di portare l'Archivio Web ad una scala adeguata, una scala-Web", ha detto Helen Hockx-Yu, Web

Archiving Programme Manager, The British Library. "IBM può aiutarci ad analizzare l'archivio web, contenente milioni di pagine, e a portare in superficie una conoscenza che, altrimenti, sarebbe molto difficile da scoprire con i metodi di ricerca tradizionali."

Che si tratti di una persona interessata al proprio albero genealogico o di uno studente al lavoro su un progetto per la scuola, le persone hanno bisogno di aiuto nell'orientarsi in questo oceano di informazioni su Web, in continuo aumento. Per esempio, le elezioni del 2005 hanno visto il primo tentativo, da parte dei politici inglesi, di usare il web come strumento di campagna politica. Si prevede che questo utilizzo del web avrà un'esplosione per le elezioni del 2010, e i dati raccolti nel 2005 consentiranno ai ricercatori di studiare l'evoluzione del rapporto tra politica e web per accedere ad una fonte primaria di materiali di grande valore.

BigSheets: i fondamenti tecnici

Quest'anno, la mole di informazioni digitali si stima raggiungerà i 988 esabite, l'equivalente di una catena di libri dal Sole a Plutone e ritorno. Il Web sta esplodendo di informazioni ed i business professionals vogliono accedere a quelle informazioni – strutturate e non strutturate – per avere una visione più approfondita del loro lavoro. IBM BigSheets è un motore di approfondimento che aiuta le aziende a ottenere una vista migliore da data set di grandi dimensioni ed in breve tempo. Costruito sul framework di Apache Hadoop, IBM BigSheets è capace di elaborare grandi quantità di dati velocemente ed in modo efficace.

IBM BigSheets è il prototipo di una nuova tecnologia. Gli utenti possono esplorare e generare nuovi approfondimenti dei dati utilizzando un'applicazione Web e poi il software IBM pubblica data feeds secondo gli standard Web 2.0, ricercabili dai clienti della British Library.

BigSheets è l'estensione del paradigma del mashup, che integra gigabytes, terabytes, o petabytes di dati non strutturati provenienti da repository su Web; raccoglie una grande quantità di dati non strutturati partendo da user-defined seed URLs; estrae ed arricchisce l'informazione usando un'architettura di gestione delle informazioni non strutturate; e permette all'utente di esplorare e visualizzare queste informazioni in un contesto specifico, ritagliato per l'utente. Per esempio, gli utenti possono vedere i risultati delle ricerche in un grafico a torta e vedere i dati in una tag cloud.

Per maggiori informazioni sui progetti di IBM nel campo della Emerging Technology, visitate il sito <http://www.ibm.com/software/ebusiness/jstart/>

Per maggiori informazioni sul British Library Web Archiving Programme, visitate <http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html>