Utilizzo di basi dati bibliografiche per analisi fattuali

Un'esperienza di data mining *e* data building *da una bibliografia di articoli di periodici* business oriented

di Marco Giarratana e Piero Cavaleri

Il progressivo ampliarsi delle basi dati di citazioni di articoli di periodici, il loro arricchirsi di indici semantici, di abstract e di testi completi, consente di ipotizzare utilizzi di queste risorse che vanno al di là della semplice ricerca e consultazione degli articoli.

In particolare le basi dati orientate al management e al business possono consentire interessanti operazioni di *data mining*[1] e *data building* per la creazione di nuove informazioni sul comportamento strategico delle aziende.

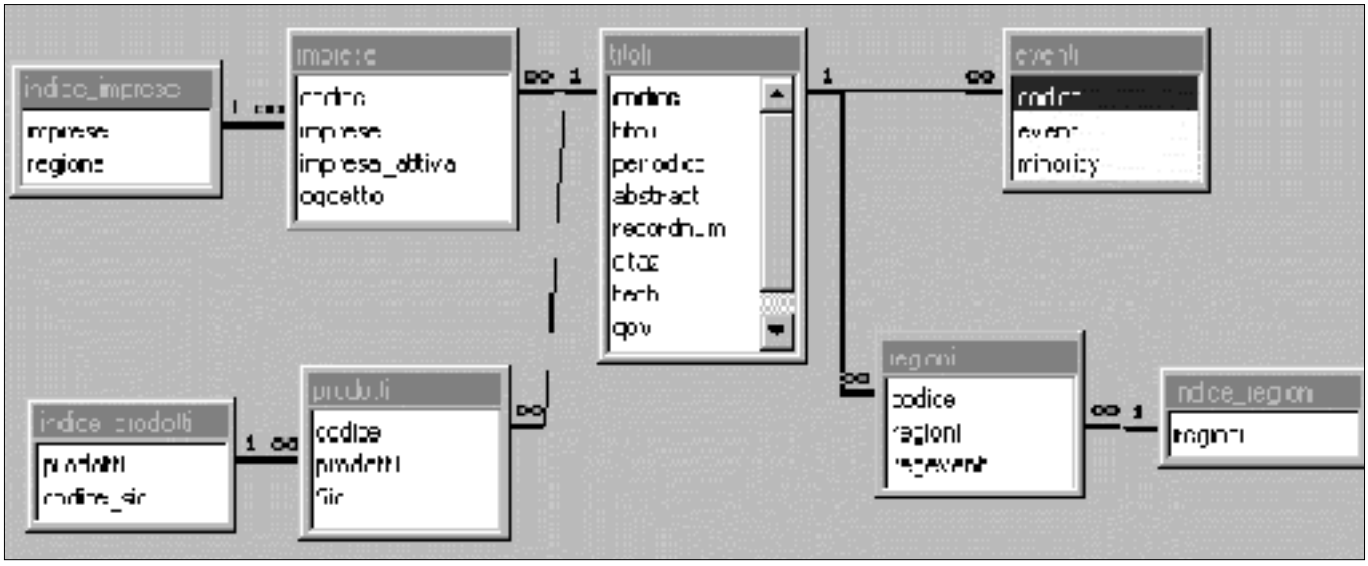
Nell'ambito di un progetto finanziato dall'Unione europea (progetto TSER "Dynacom"), un gruppo di ricercatori del Libero istituto universitario C. Cattaneo di Castellanza e dell'Università di Urbino, coordinati da Salvatore Torrisi, ha sperimentato una tale operazione utilizzando la base dati Promt Online (<http://www.insitepro.com>) di Information Access Company, scelta ed acquisita con la fattiva collaborazione della biblioteca del LIUC e di Cenfor International, che ha gestito il rapporto con il produttore.

Promt Online permette la ricerca e la consultazione degli articoli (o degli abstract) di oltre 1.000 periodici e riviste di natura economica, finanziaria e di business. Gli articoli sono indicizzati per imprese citate, prodotti, tipologie di eventi e paesi coinvolti. I prodotti sono classificati in base alla Standard Industrial Classification (SIC); gli eventi in base ad una classificazione gerarchica definita da IAC (p. es. *38 "Licensee and sales agreements", 380 "Strategic alliances", 389 "Partner-ships"*).

Scopo del lavoro è stato, partendo da un archivio tipicamente bibliografico, la creazione di un database da cui fosse possibile evidenziare le principali strategie adottate in specifici settori ad alta tecnologia. Il gruppo di ricerca ha selezionato l'insieme degli articoli relativi ad un campione di imprese elettroniche e chimico-farmaceutiche per il periodo 1993-1998, massima estensione temporale del database.

Utilizzando come base le segnalazioni di articoli contenute in Promt Online, l'obiettivo è di alimentare un archivio evento-centrico[2] che contenga i dati relativi a tutti gli eventi interessanti ai fini dell'analisi e degli studi successivi. Per evento si deve intendere ogni operazione di investimento, disinvestimento, alleanza, cooperazione o acquisto che coinvolga una delle imprese del campione durante il periodo

Figura 1. Schema del database iniziale



temporale considerato.
Al fine di trasformare le segnalazioni bibliografiche in dati sugli eventi è stato necessario apportare una serie di elaborazioni, automatiche e manuali, qui di seguito descritte nella loro essenzialità.
La presenza in ogni record selezionato del testo dell'articolo o di un abstract esauriente, ha consentito di ricostruire o di specificare informazioni che in base ai soli indici risultavano ambigue o incomplete.

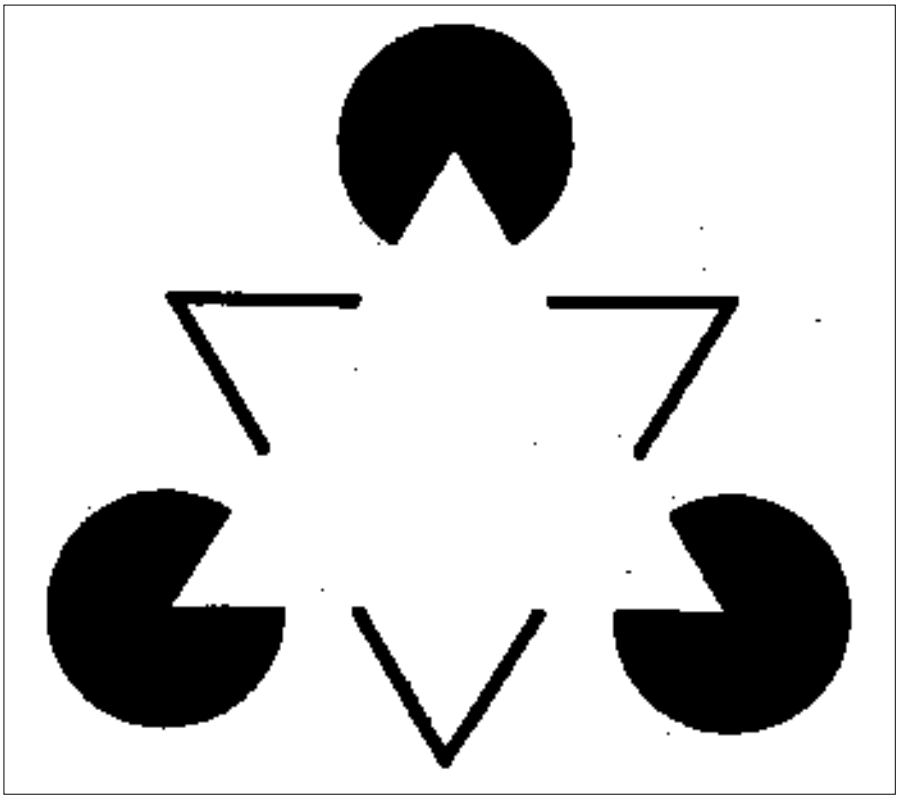
Tavola 1. Scheda estratta da Promt Online

Companies
Intel Corp.
iCat Corp.
Product Codes & Names
Semiconductor Devices (3674000)
Computer Software (7372000)
Event Codes & Names
Acquisitions & mergers (150)
Asset sales & divestitures (160)
Geographic Codes & Names
United States (1USA)
Publication Information Acquisition: Intel to Acquire iCat.
(Company Business and Marketing) (Brief Article)
EDGE: Work-Group Computing Report, Dec 7, 1998, pNA
Full Text
Intel Corporation announced Monday that it has entered into a definitive agreement to acquire the assets of iCat Corporation. The transaction will be made through a subsidiary of Intel and is expected to be completed before the end of 1998. Privately held i Cat is a leading provider of e-commerce software and services. The software aids in the creation of Web-based storefronts that include secure transaction processing capabilities. iCat Corporation, founded in 1993 and based in Seattle, provides corporations and merchants worldwide with the information, software, and services they need to be successful in e-commerce. iCat's award-winning, entry-level e-commerce solutions are used by thousands of small and mid-size businesses. Additional information about iCat is available at www.icat.com. Intel, the world's largest chip maker, is also a leading manufacturer of computer, networking and communications products. Additional about Intel is available at www.intel.com.
Full Text COPYRIGHT 1998 EDGE Publiching
Record Number: BP5336734

Elaborazioni compiute sui dati

Importazione delle notizie bibliografiche in un database relazionale

Le notizie bibliografiche individuate come rilevanti in Promt Online sono state salvate in forma di schede, finalizzate alla visualizzazione o alla stampa, prive di una vera e propria struttura automaticamente identificabile (vedi Tavola 1); per poter ottenere un archivio utilizzabile per i fini della ricerca le singole schede sono state rielaborate, utilizzando un programma di conversione scritto appositamente, e i dati risultanti sono stati memorizzati in un database relazionale (vedi Figura 1). ►



L'assenza di un formato di esportazione dei record orientato ad ulteriori elaborazioni ha reso questa fase particolarmente complessa e laboriosa.

Computo delle citazioni di un medesimo evento

Uno stesso evento (ad esempio la creazione di una data *joint venture*) può essere citato in vari articoli, di periodici diversi o, in casi più rari, del medesimo periodico.
Poiché la finalità della ricerca richiede di costruire un database finale evento-centrico, si è dovuto procedere ad eliminare tutti gli articoli che citino un evento già individuato.
Il numero delle citazioni di ogni evento è stato comunque rilevato onde non perdere un dato significativo per valutare l'importanza relativa dell'evento stesso.
Ad un primo esame è possibile notare come, nonostante l'insieme dei periodici inclusi in Promt Online sia chiaramente distorto in favore di quelli americani, ciò sembra riflettersi esclusivamente sulla numerosità delle citazioni, senza diminuire la completezza dell'informazione sugli eventi.
Normalmente, gli eventi che coinvolgono imprese americane hanno un numero di citazioni, maggiore di quelli che riguardano esclusivamente imprese europee e asiatiche, ma non si riscontra alcuna distorsione rilevante nella numerosità degli eventi coinvolgenti imprese americane, europee e asiatiche.

Individuazione univoca della tipologia di evento

Gli articoli sono spesso indicizzati con più categorie di eventi, mentre, per i fini della ricerca in corso, risulta fondamentale evidenziare in modo univoco la categoria prevalente per lo specifico evento cui l'articolo fa riferimento.
Ad esempio se l'impresa A acquisisce l'impresa B da C, saranno specificati eventi quali *"Acquisition&Merger"*, poiché A acquista, *"Asset sales & divestitures"*, perché C vende, *"Use of founds"*, perché c'è un trasferimento di fondi.
Al fine di assegnare la categoria prevalente in modo uniforme sono state definite delle linee guida per ognuna delle combinazioni rilevate. Naturalmente le linee guida sono state definite solo per le combinazioni di categorie di eventi rilevate in un numero ampio di casi, mentre valutazioni puntuali sono state effettuate per le combinazioni sporadiche.
Il caso di cui al paragrafo precedente è stato costantemente risolto con l'assegnazione del descrittore *"Acquisition&Merger"*.

Indicazione di una relazione causale

Tra le imprese segnalate è stato necessario individuare quale fosse il soggetto attivo, quale il soggetto passivo e quale, eventualmente, l'oggetto. Per alcune tipologie di eventi come *joint venture* o *strategic alliance*, in cui le imprese si trovano sullo stesso piano, questa distinzione non ha rilevanza, ma per altre categorie è fondamentale. Nell'esempio sopracitato delle tre imprese A, B e C è necessario sapere con precisione quale impresa ha acquistato, quale impresa ha venduto e quale impresa è stata l'oggetto della transazione.

Specificazione della localizzazione dell'evento e della nazionalità delle imprese

L'indicizzazione per luoghi geografici fornita da IAC si riferisce all'articolo nel suo complesso senza alcuna distinzione di quali termini si riferiscano alle singole imprese coinvolte e di quali diano conto del luogo in cui una data operazio-

ne è avvenuta. Anche in questo caso è stato necessario fornire una corrispondenza univoca tra la nazionalità dell'evento e la nazionalità delle imprese coinvolte. Ad esempio, se IBM e Sony creano una *joint venture* chiamata "Zeta" in Italia, il record estratto da Promt Online conterrà nel campo *"Country"*: USA, Giappone e Italia. In questo caso è necessario procedere ad identificare le relazioni impresa-paese ed evento-paese creando gli opportuni legami: USA sarà associato ad IBM, Giappone a Sony, Italia all'evento e all'impresa "Zeta".

Settore tecnologico principale

In Promt Online sono assegnati ad ogni articolo indici SIC relativi al settore del business principale delle imprese coinvolte e, nei casi di non coincidenza o di possibilità di maggior specificazione, ai prodotti eventualmente oggetto della specifica transazione. La presenza di indici SIC relativi genericamente alle imprese ma privi di rilevanza riguardo la specifica transazione non era coerente con i fini della ricerca. È stato quindi necessario individuare univocamente quale fosse il codice SIC in grado di descrivere il settore tecnologico dell'evento.
Nel caso IBM (il cui *core business* è indicato con il codice *3570 "Office and computing machine"*) abbia fornito a Boeing (core business *3720 "Aircraft"*) software (codice *7372 "Software"*) l'evento verrà individuato come relativo al settore del software, per cui il codice 7372 sarà quello prescelto.
L'insieme delle elaborazioni e dell'analisi dà come risultato quanto esposto nella Tavola 2.

Caratteristiche del database finale

Il database ottenuto attraverso queste operazioni permetterà, in primo luogo, di evidenziare le strategie di ogni impresa del campione secondo criteri tecnologici, geografici e di tipologia di operazioni, il tutto inserito in un contesto temporale.
Sarà infatti possibile studiare in quali settori ogni singola impresa ha investito, in quali ha invece disinvestito, le forme giuridiche utilizzate (acquisizioni, *joint venture*, accordi strategici) e quale sia stata la diversificazione geografica, tecnologica e delle linee di business.
Sarà possibile disegnare la "mappa" della rete di partner che ogni impresa ha costituito, individuando le imprese con cui più frequentemente ha concluso accordi commerciali e finanziari, ha sviluppato prodotti o ha finanziato attività di ricerca tecnologica.
Oltre a rappresentare le strategie delle imprese di partenza, il database consentirà di individuare, in base alla partecipazione ad un numero consistente di eventi, eventuali imprese estranee al campione che stiano assumendo un ruolo rilevante nel settore, pur non avendo dimensioni tali da collocarle in posizioni di primo piano per vendite o numero di addetti.
Un'altra caratteristica significativa è il numero di citazioni relative ad ogni evento, che permette di inferire il rilievo dell'evento stesso, integrando in tal modo l'informazione relativa al numero degli eventi, al fine di capire il reale impegno delle imprese nei singoli settori.
È possibile notare come il prodotto finito costruito e pensato per fini puramente accademici possa avere anche un interessante utilizzo a livello privato di business.
Se corrisponde a verità che le imprese, specialmente quelle attive in settori ad alta turbolenza tecnologica e competitiva, sono sempre alla ricerca di un vantaggio competitivo da poter sfruttare, allora l'effettiva possibilità di possedere le informazioni strategiche necessarie rappresenta una chiave importante per la formazione di una possibile superiorità competitiva, intendendo per informazione strategica i *patterns* lungo cui si evolve un settore o le scelte strategiche, siano esse vincenti o meno, attuate dai propri diretti concorrenti.
Un tipo di database così strutturato può ben quindi rappresentare un importante strumento per guidare il business di un'impresa o per meglio monitorare le proprie scelte all'interno di un contesto strategico globale.
Considerata la rilevanza per le imprese della definizione di corrette strategie, il *data mining*, in generale ed in particolare su archivi del tipo qui considerato, non dovrebbe essere pensato come uno strumento utile per poter meglio fare business, ma come qualcosa di non scindibile dal business stesso. Un'impresa troverebbe molto arduo attuare delle strategie efficaci nel lungo periodo senza acqui- ➤

Tavola 2. Informazioni ottenute a partire dalla scheda della tavola 1

Impresa Attiva: Intel Corp.
Impresa Oggetto: iCat Corp.
Evento: Acquisition (15)
Regione Impresa Attiva: USA
Regione Impresa Oggetto: USA
Regione Evento: USA
Codice SIC: Computer Software (7372000)
N. Citazioni: 6
Anno: 1998

sire ed elaborare, in base a criteri o chiavi di lettura sviluppate da economisti e/o esperti di settore, il maggior numero di informazioni disponibili.

Conclusioni

Il problema fondamentale connesso al valore economico dell'informazione non è da ricercarsi dal lato quantitativo, inteso come numerosità delle informazioni possedute, ma qualitativo; a riguardo ci riferiamo alla efficienza con cui le informazioni sono raccolte, elaborate, utilizzate.
Soprattutto a causa dell'avvento della nuova tecnologia elettronica, sempre più evoluta, si è potuto sperimentare una letterale esplosione delle informazioni disponibili. Il gruppo di ricerca ha notato una differenza molto rilevante, da questo punto di vista, tra i database on line o su cd-rom, attualmente utilizzati, e le fonti cartacee già impiegate per precedenti ricerche.
L'offerta di più informazioni, pur essendo un fattore di miglioramento estremamente importante al fine di ottenere una visione più aderente ai reali comportamenti delle imprese, pone un problema di ridondanza e di rilevanza. Questa crescita dal punto di vista quantitativo delle informazioni ha reso necessario impostare tutto un piano di lavoro teso a diminuire la ripetitività delle informazioni, accrescendone in modo significativo la qualità. A tale riguardo, per fornire un'idea di come la riduzione dal punto di vista quantitativo sia stata significativa, pare sufficiente far notare come l'insieme grezzo di più di 40.000 articoli sia stato collegato alla fine a circa 15.000 eventi.
Questa riduzione è andata di pari passo alla specificazione delle informazioni disponibili sotto forma di indici; il database finale si è così venuto a trasformare in una fonte informativa immediatamente interrogabile e consultabile che fornisce una visione di insieme delle strategie delle maggiori imprese elettroniche e chimico-farmaceutiche mondiali.
Il cuore del lavoro è rappresentato quindi da un processo in cui si sono trasformati una serie di articoli, che rappresentano i puri dati grezzi, in informazione facilmente utilizzabile. Dal punto di vista accademico e della ricerca si sta costruendo, a partire da informazioni strutturate per l'*information retreival*, una base dati fondamentalmente numerica, pronta per applicazioni di natura econometrica e statistica.
Risulta fondamentale ricordare l'importanza della definizione chiara ed esplicita, oltre che degli obbiettivi della ricerca, di una metodologia di "elaborazione" dei dati, in cui siano stabiliti, fino dal principio, le tipologie di dati che si vogliono ottenere, la struttura del database finale, le regole basilari su cui operare, le tecniche da applicare. È un errore sostanziale nella creazione di un database adottare l'approccio *"go as fast as you can and don't look for a map"*.
La metodologia di lavoro è un fattore critico per la buona riuscita di un lavoro di *data building* e *data mining*, un fattore spesso ignorato, poiché molte volte è più deleterio avere un metodo di costruzione di un database non idoneo che non averne affatto.
Un altro fattore determinante, specie rispetto alla coerenza delle scelte e al contenimento dei tempi, è stata la divisione del lavoro tra i vari membri del gruppo di ricerca, in modo da ottenere ampie economie di specializzazione nella fase di pulitura del database, assegnando poi ad uno specifico soggetto la delicata fase di aggregazione e controllo delle varie parti in cui il lavoro risulta scomposto.
La scelta di costruire un database relazionale ha consentito di ottenere un prodotto molto flessibile, facilmente modificabile e aggiornabile; il database potrà, inoltre, essere unito e messo in relazione con altri database che forniscono, ad esempio, i dati riguardanti la struttura e i risultati economici delle imprese, gli andamenti nei diversi settori tecnologici e gli indicatori fondamentali delle economie delle nazioni delle imprese presenti nel campione.
Questa esperienza mostra che un utilizzo delle basi di dati bibliografiche come fonti di informazioni di tipo fattuale, anche al di fuori del campo della misurazione della rilevanza scientifica dei documenti, non solo è possibile ma può dare risultati rilevanti. Le biblioteche, in particolare quelle universitarie, che investono sempre più risorse per acquisire l'accesso a queste basi dati, dovranno compiere le proprie scelte valutando attentamente la possibilità di impieghi di questo tipo.
Questo comporterà sicuramente la crescita di capacità e conoscenze da parte di tutti i soggetti coinvolti: produttori, intermediari, bibliotecari e utilizzatori finali. In particolare, i produttori dovrebbero rendere esplicite alcune informazioni, già elaborate concettualmente al momento dell'indicizzazione dei documenti, e prevedere la possibilità di esportare dei dati finalizzata a successive elaborazioni.
Vorremmo inoltre sottolineare che la garanzia del mantenimento dell'accesso alle informazioni, omogeneamente strutturate e indicizzate, relative ad un arco temporale più ampio possibile è di estremo rilievo, poiché l'analisi longitudinale dei fenomeni considerati è fondamentale in questo tipo di utilizzo dei database bibliografici.
Gli intermediari, agenti che si specializzano nella distribuzione di prodotti elettronici, potranno svolgere un ruolo significativo se sa-

pranno indicare non solo quali database forniscono informazioni bibliografiche in determinati settori, ma anche quali rappresentano una valida fonte di dati da elaborare per specifiche finalità informative.
Per i bibliotecari si pone il problema di ampliare la conoscenza delle basi dati bibliografiche, considerando non solo le necessità attuali degli utenti di accesso ai documenti, ma anche quali informazioni primarie siano potenzialmente deducibili dai dati in esse contenuti. Un ruolo importante i bibliotecari potranno svolgerlo riguardo al trattamento automatico dei dati "scaricati" da questi archivi, il cui uso quotidiano e continuativo è esclusivamente di loro competenza.
La diffusione della conoscenza di queste potenzialità tra gli utilizzatori finali è importante, specie in ambito universitario, anche in funzione della valutazione degli investimenti che le istituzioni di appartenenza compiono per acquisire queste fonti informative. ■

Note

[1] L'operazione di *data mining* descritta in questo articolo ha delle caratteristiche inusuali essendo riferita all'utilizzo di informazioni pubbliche e non come più spesso avviene di archivi aziendali. Riguardo l'importanza del *data mining* in generale per la gestione delle imprese si vedano:
D. ASBRAND, *Making money from data*, "Datamation", November 1998 (http://www.datamation.com/PlugIn/issues/1998/november/11tran.html); K. WATTERSON, *When it comes to choosing a database*, the object is value, "Datamation", 44 (1998), 1, p. 100-106. *Cost-cutting activity*, "The Economist", 348 (1998), 8079, August, 1st-7st, p. 61.
[2] La struttura del database in cui sono importati i dati ha come tabella principale quella degli articoli (Titoli). Questa tabella, dopo l'eliminazione delle segnalazioni relative al medesime evento e l'individuazione del tipo di evento principale, contiene uno e un solo record per ogni evento, cioè per ogni operazione intrapresa da una azienda del campione, e si fonde con quella dei tipi di eventi (Eventi) con cui viene ad essere in relazione biunivoca. Ad essa vengono collegate le tabelle relative alle imprese, ai settori tecnologici, ai paesi.