In questo modulo sarà presentata un'introduzione ai concetti teorici di del Geographic information retrieval, per la cui applicazione si rimanda alla letteratura corrente.

GEOGRAPHIC INFORMATION RETRIEVAL

Prof. Michele Tarantino

Il Geographic Information Retrieval è l'insieme di tutti i metodi, algoritmi e modelli per il reperimento delle informazioni di carattere spaziale, che vengono identificate da un'apposita geometria. Nell'information retrieval classico, i documenti reperiti vengono selezionati in base ai termini che sono presenti nell'interrogazione, invece in un motore di ricerca con capacità geografiche, deve essere dato un preciso significato ai termini geografici. Non trattandosi di informazione di tipo testuale, bensì di tipo spaziale e/o topologico, il problema di selezionare un'informazione geografica, può essere ricondotto all'elaborazione di similarità tra due località geografiche, una associata con l'interrogazione e l'altra con il documento: questo confronto avviene per mezzo di relazioni metriche (definiscono la distanza tra oggetti), direzionali (*più a est, a nord di*,…) e topologiche (*adiacente a, contiene,…*).

Interrogazioni spaziali

Ogni documento spaziale (o geografico) contiene informazione di tipo spaziale e informazione di tipo testuale. Pertanto, un interrogazione di un sistema GIR deve specificare la parte di informazione testuale, la parte di informazione spaziale e la relazione spaziale. Quindi, ogni interrogazione di un sistema GIR può essere espressa mediante una tripla:

<Cosa, Relazione, Dove>

dove "*Cosa*" identifica il termine utilizzato per specificare un'informazione non spaziale, "*Dove*" identifica un'unica area geografica di interesse in cui deve essere effettuata la ricerca e "*Relazione*" specifica la relazione spaziale che intercorre tra i connettivi "*Cosa*" e "*Dove*". Chiaramente, interrogazioni più complesse possono richiedere più vincoli sulle relazioni spaziali: in tal caso l'interrogazione sarà costituita da un insieme specifico di triple.

Come per i classici sistemi di information retrieval, anche per quelli geografici si può definire una funzione di similitudine che calcola il peso di un'interrogazione (geografica) generica con ogni documento, ed è definita nel seguente modo:

$$Sim\ (q, d) = b \times TextSim\ (T_q, T_d) + (1 - b) \times GeographicSim\ (S_q, S_d)$$

Dove T_q e T_d sono i testi (informazioni non spaziali) rispettivamente dell'interrogazione e del documento, S_q e S_d rappresentano rispettivamente l'informazione spaziale dell'interrogazione e del documento. *TextSim* rappresenta la funzione di similitudine per le informazioni non geografiche (testuali) e, analogamente, *GeograhicSim* rappresenta la funzione di similitudine per le informazioni di tipo geografico; *b*

è una costante che varia nell'intervallo [0,1]. Informalmente, data un'interrogazione con valori testuali e valori geografici, il sistema GIR deve recuperare tutti i documenti che sono rilevanti per l'informazione testuale e che soddisfano la relazione geografica di similitudine.

Data un'interrogazione con rappresentazione spaziale S_q e un documento con rappresentazione spaziale S_d , si definiscono le seguenti funzioni geometriche:

$$\textbf{Inclusione: } \text{Inclusion}(S_q, S_d) = \begin{cases} \dfrac{(\text{NumDiscendenze}(S_d) + 1)}{(\text{NumDiscendenze}(S_q) + 1)} & \text{Se } S_d \text{ è incluso in } S_q \\ 0 & \text{Altrimenti} \end{cases}$$

Dove S_q e S_d rappresentano rispettivamente l'informazione spaziale dell'interrogazione e del documento, *NumDiscendenze (S) + 1* definisce il numero di *prospettive spaziali* di S, ossia definisce il numero di *sottoregioni* dell'area geografica di interesse. Questa formula restituisce un valore prossimo a 1 quando le due regioni sono uguali (valore esattamente uguale a 1) o quando le due regioni sono simili.

$$\textbf{Prossimità: } \text{Proximity}(S_q, S_d) = \frac{1}{1 + \text{Distanza}(S_q, S_d) / \text{Diagonal}(S_q)}$$

Dove *Distanza* (S_q, S_d) è appunto la distanza normalizzata della diagonale del rettangolo che limita l'area geografica dell'interrogazione, ossia l'area geografica che viene delimitata da un rettangolo in cui tutti i lati sono a contatto con almeno un punto dell'area spaziale considerata.

$$\textbf{Siblings: } \text{Siblings}(S_q, S_d) = \begin{cases} 1 & \text{Se esiste } S_x \text{ tale che parent}(S_q) = S_x \\ & \text{\& parent}(S_d) = S_x \\ 0 & \text{Altrimenti} \end{cases}$$

La funzione binaria *Siblings* (S_q, S_d) definisce se le due regioni S_q e S_d sono sottoregioni, incluse nella stessa sottoregione, mediante la funzione parent (S_x).

Avendo definito queste tre funzioni spaziali fondamentali la funzione di similitudine geografica descritta precedentemente, può essere formulata nel seguente modo:

$$GeographicSim\ (S_q, S_d) = bb\ x\ (Inclusion\ (S_q, S_d) + Proximity\ (S_q, S_d)) + (1 - bb)\ x\ Siblings\ (S_q, S_d)$$

Dove bb è una costante che appartiene all'intervallo [0,1] e *Inclusion* (S_q, S_d), *Proximity* (S_q, S_d) e *Siblings* (S_q, S_d) sono le funzioni descritte sopra.

Rappresentazione dello spazio geografico

Nei sistemi di Geographic Information Retrieval si definisce un'ontologia di diverse forme geometriche, per rappresentare l'intera superficie del pianeta. Alcuni di questi spazi possono essere usati per i documenti che presentano informazione spaziale, ma che non sono associati con nessuna forma geometrica presente nell'ontologia. In alcuni casi, la forma assegnata ad una superficie può essere definita come interpolazione dei punti mediani di tutte le corrispondenti forme geometriche. In altri casi, invece, la superficie viene definita mediante un *rettangolo di limite minimo* (***M**inimum **B**ounding **R**ectangles* – ***MBR**s*). Una soluzione al problema di rappresentazione delle superfici, deriva dall'utilizzo del MBR. Fissata una forma geometrica non presente nell'ontologia, si definisce il rettangolo di limite minimo che contiene tutta la forma geometrica considerata, in cui almeno un punto di ogni lato del rettangolo è a contatto con il confine della figura geometrica considerata. Per una maggiore comprensione si consideri la Figura 1:

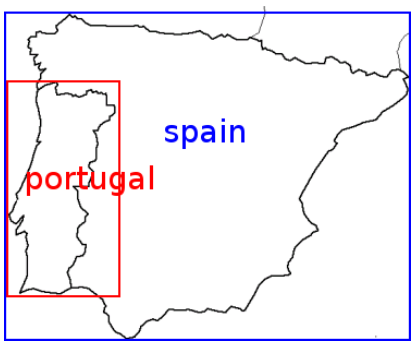

Figura 1: Esempio di utilizzo del rettangolo di limite minimo

Si nota che il Portogallo è rappresentato da una figura geometrica che non appartiene all'ontologia del Sistema GIR. Utilizzando il metodo del triangolo di limite minimo, si riesce a delimitare l'area del Portogallo con una figura geometrica presente nell'ontologia. Così facendo, però, parte della superficie della Spagna viene inglobata all'interno del rettangolo, e ad esempio tutte le città che si trovano in questo triangolo vengono definite appartenenti al Portogallo, quando in realtà questo non è vero. Quindi, si possono creare

problemi di rappresentazione delle informazioni spaziali e in particolar modo i problemi più rilevanti sono: le caratteristiche del dominio spaziale, le relazioni spaziali che intercorrono tra gli oggetti e il trattamento di informazioni incomplete (ad esempio confini indeterminati e misure imprecise).

Per risolvere in parte questo problema si possono definire tre tipi di differenti geometrie per la rappresentazione delle informazioni spaziali:

- poligoni: per rappresentare continenti, province, regioni,…
- linee: per rappresentare fiumi, confini,…
- Punti: per rappresentare città, luoghi precisi,…

Esempi di interrogazioni Geografiche

Fino adesso si sono descritti i criteri generali che un sistema di Geographic Information Retrieval deve avere, ma quali sono effettivamente le applicazioni di un tale sistema? In linea di massima, con un sistema di Geographic Information Retrieval si possono recuperare tutte le informazioni di carattere spaziale (quindi anche visualizzabili come immagine) che soddisfano determinati criteri. Un esempio di interrogazione spaziale è la seguente:

Text = "Restaurant" AND GeograpghicSim (Danube): restituisce tutti i documenti che contengono informazioni relative ai ristoranti, e la cui geografia si trova in prossimità del Danubio, ottenendo i risultati riportati nella tabella 1:

Rank	Assigned scope	Text	Geo	Final
1	Wien	0.795	0.473	0.505
2	Bayern	0.411	0.420	0.419
3	Strasbourg	0.715	0.299	0.341
4	Austria	0.639	0.290	0.325
5	Brno	0.804	0.263	0.317
6	Vienna	0.421	0.290	0.303
7	Austria	0.370	0.290	0.298
8	Bulgaria	0.580	0.266	0.298

Tabella 1: Risultati dell'interrogazione Text = "Restaurant" AND GeograpghicSim (Danube) senza MBR

La stessa interrogazione può essere effettuata considerando il rettangolo di limite minimo, ottenendo i risultati presentati in tabella 2:

Rank	Assigned scope	Text	Geo	Final
1	Wien	0.795	0.500	0.529
2	Brasov	0.781	0.500	0.528
3	San Marino	0.764	0.500	0.526
4	Switzerland	0.756	0.500	0.525
5	Berne	0.734	0.500	0.523
6	Monza	0.716	0.500	0.521
7	Pisa	0.700	0.500	0.520
8	Liguria	0.674	0.500	0.517

Tabella 2: Risultati dell'interrogazione Text = "Restaurant" AND GeograpghicSim (Danube) con MBR

Dai risultati riportati dagli esperimenti con *GeoCLEF 2005* e riportati in [Andrade, Silva] si può notare come i risultati stessi variano a seconda del tipo di geometria considerata. Considerando l'area come MBR, sono reperiti i documenti che hanno un'informazione spaziale differente dall'area in cui si voleva esattamente effettuare la ricerca. La causa di questa bassa prestazione con MBR è dovuta al fatto che i termini geografici vengo "strappati" dall'interrogazione originale, e inoltre dato che l'interrogazione è stata fatta su un termine geografico "molto esteso" è chiaro che i risultati non sono ottimali. Quindi, alcune interrogazioni sono "più geografiche" di altre, o meglio si prestano meglio al sistema di Geographic Information Retrieval considerato.

Web Geographic Retrieval *[da completare e rivedere]*

Fino ad ora si è discusso su come vengono rappresentate le interrogazioni e quali sono i problemi relativi alla rappresentazione delle informazioni spaziali. Ma come vengono visualizzate ad esempio su web le informazioni di carattere spaziale e come vengono recuperate? Si può presentare una visione globale della struttura per il reperimento delle informazioni geografiche come in Figura 2:

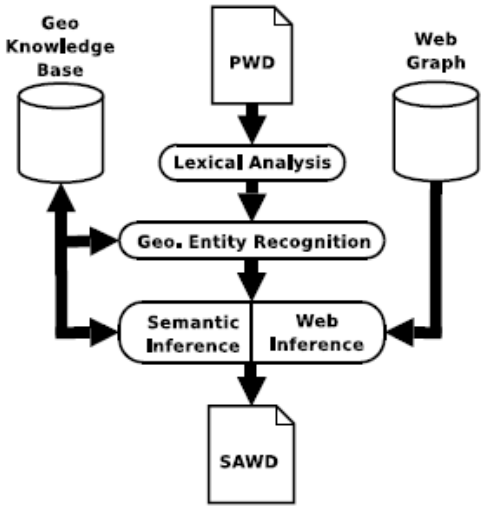

Figura 2: Tipica struttura di un Geographic Information Retrieval

La tipica struttura comprende otto elementi base di seguito descritti:

- Geographic Knowledge Base: rappresenta una base di dati che identifica le associazioni tra le entità georeferenziate (o entità spaziali) contenute nelle pagine web e le geometrie contenute nell'ontologia.
- Lexical Analysis: il testo contenuto nei documenti Web è diviso in simboli (token) che comprendono parole e/o marcatori che specificano l'informazione (tag), e questo elemento deve distinguere la sintassi testuale dalla sintassi geografica.
- Geographic Entity Recognition: le entità georeferenziate sono identificate da specifici modelli presenti nella base di dati di conoscenza geografica.
- Web Inference: elemento che propaga la "conoscenza" geografica. Ad esempio se è presente un collegamento ipertestuale da un documento ad un altro, significa che i due documenti sono correlati e quindi, se la "conoscenza" geografica del primo documento era conosciuta a priori, si può supporre che il documento a cui è collegato, abbia la stessa conoscenza geografica del primo.
- Semantic Inference: elemento che determina secondo modelli probabilistici, e in base alla conoscenza geografica, la semantica dei termini geografici.
- PWD (Purified Web Documents): elemento che raccoglie il contenuto dei vari documenti e li inserisce in uno schema comune.
- SAWD (Scope Augmented Web Documents): elemento che raccoglie la conoscenza geografica associata a risorse addizionali (ad esempio RDF).

Bibliografia

[Andrade, Silva] Leonardo Andrade, Mário J. Silva. Relevance Ranking for Geograhic IR.

[Chrisment, Dousset, Karouach, Mothe] Claude Chrisment, Bernard Dousset, Said Karouach, Josiane Mothe. Information Mining: extracting, exploring and visualising geo-referenced information. Workshop on Geographic Information Retrieval, Sheffield.

[Larson, Frontiera] Ray. R. Larson, Patricia Frontiera. Ranking and Representation for Geographic Information Retrieval. Proceedings of the Research and Advanced Technology. 2004.

[Sanderson, Kohler] Mark Sanderson, Janet Kohler. Analyzing geographic queries. SIGIR Workshop on Geographic Information Retrieval. 2004.

[Schilder, Versley, Habel] Frank Schilder, Yannick Versley, Christopher Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. Workshop on Geographic Information Retrieval at SIGIR. 2004.

[Silva, Martins, Chaves, Cardoso, Afonso] Mário J. Silva, Bruno Martins, Marcirio Chaves, Nuno Cardoso, Ana Paula Afonso. Adding Geographic Scope to Web Resources. CEUS-Computers, Environment and Urban Systems. 2006.

Resta connesso e informato sui prossimi eventi, corsi e seminari:

Web
www.profmicheletarantino.com

Email
profmicheletarantino@gmail.com

Telefono
349 83 54 521

Facebook
@micheletarantinodocente

Instagram
@profmicheletarantino

Hai bisogno di un modulo personalizzato? Non esitare a contattarmi!