MapPapers 2-II, 2012, pp.39-44
doi:10.4456/MAPPA.2012.03

Sulle due questioni principali inerenti l'applicazione dei modelli page rank per la determinazione del potenziale archeologico

Bini D., Dubbini N., Steffè S.

Il presente report tratta la modellazione matematica dei due principali problemi implementativi riguardanti l'applicazione di algoritmi di tipo page rank alla determinazione del potenziale archeologico: un'efficace categorizzazione dei ritrovamenti e la costruzione della matrice dei pesi. Per quanto riguarda la categorizzazione, ogni ritrovamento è descritto, dal dettaglio verso una progressiva astrazione, mediante quattro livelli di sintesi, basandosi poi su questi livelli per la definizione delle categorie. Per quanto riguarda la matrice dei pesi, si costruisce sulla base di una codifica di proprietà spaziali e di relazioni funzionali tra ritrovamenti, matematicamente definita. I passi successivi riguarderanno l'applicazione pratica di queste informazioni per produrre una sola matrice dei pesi, e per ottimizzare attraverso alcune prove, tra tutte le diverse possibilità, l'algoritmo di page rank.

Keywords: page rank, teoria della forma, categorizzazione dei ritrovamenti, relazioni funzionali tra ritrovamenti

1. Introduzione

Nel report precedente (Bini, 2011) gli autori hanno identificato nei modelli di tipo page rank le caratteristiche adatte per determinare il potenziale archeologico di una zona urbana. Questi modelli funzionano soprattutto perché permettono di codificare in termini matematici le relazioni tra i vari ritrovamenti, sia in termini spaziali (cioè che abbiano a che fare con la posizione nello spazio), sia in termini funzionali (cioè che abbiano a che fare con la funzione dei ritrovamenti). Queste relazioni sono proprio la chiave della determinazione pratica del potenziale da parte degli archeologi.

Quando l'informazione archeologica è a disposizione, due tipi di quantità sono associati nel modello page rank ad ogni ritrovamento (o più precisamente a ciascuna cella 3-dimensionale in cui viene idealmente diviso il sottosuolo): un valore che ne rappresenta l'importanza e un insieme di valori che rappresentano la 'forza' con cui la cella influenza il potenziale archeologico delle altre celle. L'idea generale è di classificare i ritrovamenti e, per ciascuna categoria, assegnare alcuni parametri che forniscano l'importanza del ritrovamento e la geometria della distribuzione di tale importanza verso gli altri ritrovamenti. Infine, nel modello page rank tutte queste informazioni verranno integrate per calcolare il potenziale archeologico in ogni cella. Una parte fondamentale dell'implementazione del modello page rank, quindi, riguarda la categorizzazione dei ritrovamenti, e la definizione di un set appropriato di parametri che possano descrivere, per ciascuna categoria, la distribuzione dei suoi pesi.

In questo report si propone una classificazione basata su considerazioni di ordine pratico, specificatamente legate al caso di studio di un'area urbana, e su precedenti lavori in materia. Inoltre daremo una serie di parametri che, per ciascuna categoria, descriva la geometria e la distribuzione dei pesi.

Il report è organizzato come segue: nella sezione 2 presentiamo la categorizzazione dei ritrovamenti, mentre la sezione 3 riguarda la costruzione della matrice dei pesi, attraverso un modello matematico delle proprietà spaziali e delle relazioni funzionali tra ritrovamenti. Nella sezione 4 presentiamo conclusioni e futuri sviluppi. Nell'appendice è illustrata brevemente un'introduzione alla teoria matematica delle forme, utile per specificare le proprietà spaziali delle categorie di ritrovamenti.

2. Determinazione delle categorie

Una categorizzazione dei ritrovamenti è fondamentale per la creazione di un algoritmo, perché, se considerassimo ogni ritrovamento con tutte le sue peculiarità, la variabilità sarebbe troppo alta. La classificazione aiuterà, quindi, un'efficace implementazione dell'algoritmo e permetterà di descrivere la distribuzione del potenziale archeologico tra i ritrovamenti. Tuttavia, nella scelta delle categorie, due opposte esigenze devono essere prese in considerazione: primo, una categorizzazione dovrebbe essere sufficientemente generale per poter essere applicata anche in contesti diversi (e in diversi periodi archeologici, come vedremo) e non solo nel caso di studio specifico; al contrario, una categorizzazione deve essere sufficientemente dettagliata per assicurare che la geometria, la distribuzione dei pesi, e i parametri, descrivano le peculiarità di ciascuna categoria.

Per illustrare queste esigenze opposte in modo semplice, supponiamo che la classificazione sia la più dettagliata possibile, e cioè, che ogni ritrovamento abbia una propria categoria. In questo caso assegneremmo, per ciascun ritrovamento, parametri che descrivono l'importanza e la 'diffusione' del potenziale. Questo sarebbe magari un buon lavoro per il caso in esame, ma del tutto inutile per qualsiasi altra analisi, dato che ogni categoria dovrebbe di nuovo essere definita ex novo.

Al fine di definire le categorie, e di farlo con un'adeguata generalità, abbiamo seguito la procedura già adottata dagli autori per la catalogazione dei dati archeologici (Anichini, 2004). Questo metodo è il risultato di varie discussioni con il team archeologico, in cui si contrapponeva il loro bisogno di "particolarità", conferendo importanza ad ogni ritrovamento, alla nostra necessità di "generalità", per favorire un'implementazione efficace dell'algoritmo. Quindi ogni ritrovamento è stato classificato mediante 4 livelli di progressiva astrazione, che descrivono in ordine crescente di generalità la 'struttura' in cui il ritrovamento era inserito. Mentre il quarto livello, per esempio, è semplicemente descrittiva del ritrovamento, o fornisce la sua immediata interpretazione/funzione (ad esempio il ritrovamento 'finestra'), gli altri livelli descrivono la struttura in cui è inserito l'oggetto, a diversi livelli di complessità, o di funzionalità: il ritrovamento 'finestra' si trova dentro una 'domus' (il terzo livello), che a sua volta è un 'edificio residenziale' (il secondo livello), che è un caso particolare di 'area ad uso privato' (il primo livello). Il database che abbiamo creato contiene circa 250 voci del primo livello, circa 200 del secondo livello, circa 40 del terzo livello, e circa 10 del quarto livello, il più generale. Vedere (Anichini, 2012) per ulteriori dettagli.

Abbiamo scelto il terzo livello per fornire la base per la categorizzazione dei reperti, per i seguenti motivi:

- le voci del terzo livello sono le uniche a dare informazioni non solo sul luogo del ritrovamento, ma anche sui luoghi circostanti. Gli altri livelli sono o troppo specifiche o troppo generali per fornire tali informazioni. In altre parole, le voci del terzo livello consentono un'induzione spaziale sul potenziale archeologico;
- il terzo livello è portatore di un valore spaziale, funzionale, e cronologico, e contiene dunque in sé i presupposti del potenziale archeologico.

È fondamentale che le categorie siano date in modo tale che ciascuna di esse fornisca informazioni sulla zona circostante, sia per le prestazioni dell'algoritmo, sia per una determinazione accurata del potenziale archeologico. Poiché il valore del potenziale sarà assegnato per ogni cella, la possibilità di fornire informazioni sulle celle circostanti dipende anche da quanto le celle sono grandi. Non c'è nessuna generalità per questo parametro: le dimensioni 'giuste' dipendono dalla densità dei ritrovamenti, dalla vastità della zona, dalla ricchezza delle strutture e dalle fasi archeologiche, ecc. La dimensione delle celle dovrebbe essere stabilita dunque per ciascun caso di studio. Nella nostra analisi abbiamo tenuto conto di due distinti ordini di considerazioni, uno per la determinazione delle dimensioni di superficie (lunghezza e larghezza), e uno per la determinazione della profondità. Per quanto riguarda quest'ultima, periodi archeologici distinti identificano naturalmente un modo di partizionare il sottosuolo. D'altra parte sarebbe poco sensato avere in una stessa cella due periodi archeologici differenti. Quindi la scelta ovvia è stata quella di impostare la profondità delle celle in modo tale che ognuna di queste copra uno e un solo periodo archeologico. Si osservi che in questo modo la profondità delle celle può essere diversa da luogo a luogo, ma questo non è un problema dal punto di vista algoritmico.

Per quanto riguarda le dimensioni della superficie, dobbiamo far fronte a due esigenze opposte. Se le celle sono troppo piccole, allora troppe celle non conterranno ritrovamenti; in questo modo dovremmo assegnare il potenziale di molte celle con pochi valori noti: un problema critico, algoritmicamente parlando. D'altra parte, se le celle fossero troppo grandi, potremmo avere ritrovamenti (categorie) diversi in una singola cella, e la 'diffusione' del potenziale archeologico mediante i pesi sarebbe fortemen-

te ridotta. Nel caso in esame, l'area urbana di Pisa si estende su una superficie di 26 chilometri quadrati e una profondità massima di 10 metri circa. Considerando la presenza di fiumi e altre condizioni per cui alcune celle avranno potenziale nullo, abbiamo circa 130.000.000 di metri quadrati di sottosuolo a cui il potenziale archeologico deve essere assegnato. Il numero approssimativo di ritrovamenti (categorizzati) è di 2000, mentre la loro dimensione approssimativa può variare da meno di un metro a decine di metri. Tenendo conto di tutti questi indicatori, e delle considerazioni fatte in precedenza, abbiamo optato per una dimensione approssimativa di un quadrato di 1-2 metri di lato per ogni cella.

Un altro punto da discutere è se le il set di categorie debba essere lo stesso per tutti i periodi archeologici oppure no. Questo potrebbe essere un altro problema da prendere in considerazione nella scelta della generalità con la quale categorizzare i ritrovamenti. È utile avere la stessa serie di categorie per ciascun periodo archeologico, perché:

- consente una maggiore 'economia' nelle procedure algoritmiche, visto che le categorie e i parametri ad esse collegati possono essere definite una volta per tutte;
- le relazioni tra categorie, anche riguardanti diversi periodi archeologici, sono più facili da definire, ed allo stesso tempo inducono un livello di astrazione adeguato.

Nel nostro approccio si è scelto di definire la stessa serie di categorie per tutti i periodi archeologici, anche se naturalmente alcune di esse sono specifiche solo di alcuni periodi, come ad esempio le chiese, assenti prima dell'era cristiana. In ogni caso, queste categorie 'eccedenti' sono in numero esiguo.

3. La geometria e la distribuzione dei pesi per ogni cateogria

Una volta che le categorie sono state determinate e ciascun ritrovamento è stato assegnato alla categoria corrispondente, deve essere determinata la matrice dei pesi. In base a quanto stabilito nel nostro primo report (Bini, 2011) ogni categoria ha il suo 'valore assoluto' di potenziale, e il suo set di pesi associati. Avendo a disposizione i valori assoluti e la matrice dei pesi, può essere implementato l'algoritmo di page rank, fornendo il potenziale archeologico come output. Questa sezione espone le idee che stanno dietro alla costruzione della matrice dei pesi.

Principalmente due tipi di proprietà o relazioni tra categorie influenzano il potenziale archeologico: le abbiamo denominate proprietà spaziali e relazioni funzionali. Le proprietà spaziali hanno a che fare con la dislocazione nello spazio delle categorie, in modo tale che un ritrovamento in una cella implica la presenza dell'oggetto indicato dalla categoria corrispondente, ma la particolare distribuzione nello spazio può variare in dimensione e orientamento, a seconda della particolare realizzazione. Le relazioni funzionali tra categorie si riferiscono alle 'funzioni' che collegano alcune categorie ad altre: ad esempio una casa deve avere un pozzo o un giardino nelle vicinanze, e questo chiaramente influisce sulla determinazione del potenziale archeologico.

3.1 Proprietà spaziali delle categorie di ritrovamenti

Le proprietà spaziali di ogni categoria dovrebbero descrivere la dislocazione probabile della categoria nel sottosuolo. Dal momento che gli archeologi deducono la presenza di ogni categoria in un luogo mediante i ritrovamenti, e dal momento che ogni categoria può avere 'realizzazioni' di diverse forme, dimensioni e orientamento nel sottosuolo, ogni categoria deve essere identificata attraverso le sue caratteristiche peculiari, mentre la variazione delle sue caratteristiche che dipendono dalla particolare 'realizzazione' potrebbe essere codificata per mezzo di alcuni parametri, che poi saranno fissati di volta in volta.

Crediamo che la caratteristica spaziale peculiare di ogni categoria possa essere identificata con la sua forma, intesa come una descrizione geometrica della parte di spazio occupata dall'oggetto, prescindendo dalla posizione, dall'orientamento e dalle dimensioni. Esiste una teoria matematica che permette di dare definizioni formali e proprietà: la cosiddetta teoria delle forme (Kendall, 1989), di cui diamo una breve introduzione in appendice. Ad ogni modo la cosa importante è che la forma di ciascuna categoria, nel senso della definizione di cui sopra, possa essere codificata in modo opportuno come una caratteristica della categoria stessa. Inoltre la teoria permette anche di prendere in considerazione qualche incertezza, o errore, nella forma. Questo è utile, perché spesso nella pratica archeologica la forma non è completamente determinata o conosciuta a partire dai ritrovamenti.

Una volta che la forma è stata assegnata ad ogni categoria, la particolare realizzazione della categoria deve essere posizionata nel sottosuolo dando una stima della sua posizione, dell'orientamento e della dimensione. Schematicamente, la posizione corrisponde ad un parametro numerico di traslazione, l'orientamento ad un parametro numerico di rotazione, e la dimensione ad un parametro numerico di 'allungamento'. Questo insieme di parametri permette alla particolare realizzazione di ciascuna categoria di essere localizzata nel sottosuolo. Ulteriori dettagli sul modo di assegnare e codificare le forme sono dati nell'appendice.

3.2 Relazioni funzionali tra categorie

Le relazioni funzionali tra categorie dovrebbero stimare la probabilità, per ogni categoria, di avere un'altra categoria nello spazio vicino, a causa della loro occorrenza comune, o a causa della loro funzione. Un semplice esempio di una relazione funzionale è la presenza contemporanea di una casa, insieme a una strada, un pozzo, un giardino, spesso presenti

perché sono funzionali allo scopo per cui una casa è costruita. Quindi ciò a cui ci riferiamo con il termine 'relazioni funzionali' è dato da categorie che spesso si riscontrano insieme, perché sono costruite per raggiungere un obiettivo comune (come casa, pozzo, strada, e così via), che può essere pratico o di altro tipo (come una chiesa insieme ad un campanile, o a un cimitero).

Si osservi che le relazioni funzionali tra categorie possono essere spesso relazioni tra periodi archeologici diversi, che sono 'funzionali' in un senso ulteriore. Ad esempio, spesso sulle rovine di alcuni ritrovamenti afferenti ad una particolare categoria, nei periodi archeologici successivi è più probabile la presenza di particolari categorie. Questo può essere l'effetto di una 'continuità' in periodi diversi di una stessa funzione (ad esempio una chiesa che si imposta sulle rovine di un tempio).

Tenuto conto delle considerazioni precedenti, sono due le questioni principali per quanto riguarda i rapporti funzionali tra categorie. La prima concerne la stima della probabilità della presenza di una categoria, data la presenza di una prima categoria. Questo si farà creando un vettore 4-dimensionale $M \in [0,1]^{N\times T\times N\times T}$,, dove N è il numero totale delle categorie usate per l'analisi, e T è il numero dei periodi archeologici. L'elemento $M_{i,j,k,l}$ rappresenta quindi la probabilità della presenza della categoria i nel periodo archeologico j, data la presenza della categoria k nel periodo archeologico l. Questa probabilità sarà stimata con metodi diversi, a seconda del caso specifico. Ad esempio potrebbe essere stimata a partire da modi di procedere 'standard' legati ad un particolare periodo archeologico o ad un particolare luogo, o anche a partire da altro tipo di dati, per esempio storici, o da archivi.

La seconda questione riguarda il luogo in cui si trova ogni categoria. Una volta intuito che la presenza di una categoria implica in qualche modo la presenza di un'altra categoria, nello stesso periodo archeologico o no, deve essere eseguita una stima del luogo (cioè delle celle) occupato da quest'ultima categoria. Ecco perché le relazioni funzionali devono essere considerate soltanto dopo le relazioni spaziali. Ogni categoria deve a questo punto avere già una sua forma (nel senso specificato prima) e suoi parametri che descrivono la dislocazione nello spazio. In questo modo assegneremo alle celle i valori delle probabilità che contribuiranno alla formazione della matrice dei pesi nell'algoritmo di page rank.

Appendice

Diamo in questa appendice le basi della teoria matematica della forma (shape theory), come sviluppata da (Kendall, 1999), che verrà utilizzati per assegnare le proprietà spaziali delle categorie. Forse vale la pena notare che uno dei problemi pratici da cui la teoria della forma si è sviluppata era un problema sorto in archeologia. Citando (Kendall, 1989): "Così la serie di 52 pietre erette vicino Land 's End, in Cornovaglia, studiata in (Broadbent, 1980), fornisce

$\binom{52}{3} = 22.100$ triplette di pietre. Ci sono quelli che dicono vagamente che 'troppe' di queste sono 'quasi allineate', attribuendo a ciò pianificazione deliberata, mentre altri respingono queste conclusioni come ridicole. Chi ha ragione?". Quindi il problema consta di come quantificare la proprietà 'essere allineate', per le pietre di quel particolare sito archeologico.

La prima questione che si pone nella teoria della forma è come definire matematicamente la forma di un insieme di k punti non totalmente coincidenti nello spazio. Per quanto riguarda questo report, consideriamo solo il caso di spazi a 2 o 3 dimensioni, nonostante in generale la teoria consideri spazi di ogni dimensione. Come abbiamo già osservato, l'idea è quella di non considerare gli effetti delle traslazioni, del cambio di scala, e delle rotazioni. Di seguito una descrizione di come una definizione matematica della forma si ottiene:

1. 1. Si consideri un insieme di k punti nello spazio 3-dimensionale, dato come una matrice $X \in R^{3\times k}$, in cui ogni colonna è il vettore contenente le coordinate di uno dei k punti;
2. 2. Si porti il baricentro di tali punti all'origine, in modo che il parametro traslazione sarà il vettore delle coordinate del baricentro dei k punti. In formule, per ogni vettore colonna X_i, i=1,...,k;
$$X_i \to X_i - \frac{X_1 + ... + X_k}{k};$$
3. Per eliminare l'effetto dell'estensione nella definizione della forma, si consideri la quantità $L = \sqrt{\sum_{i=1}^{k} \|X_i\|^2}$, cioè la norma del vettore fatto dalle distanze tra ciascun punto X_i e l'origine. Si divida ciascun vettore colonna X_i per L in modo tale che L sia uguale a 1:
$$X \to \frac{1}{L} X = \frac{1}{\sqrt{\sum_{i=1}^{k} \|X_i\|^2}} X;$$
4. Ora il rango della matrice X è al massimo 2, e qundi possiamo moltiplicare X per quell'elemento del gruppo ortogonale O (k) che manda (0, ..., 0,1) nell'elemento di R^3 con tutte le coordinate uguali a $1/\sqrt{k}$. *X* ora ha l'ultima colonna uguale a 0, e quindi consideriamo $X \in R^{k\times 2}$;
5. Ora che X è fatto di k-1 vettori che si sommano al vettore unità, possiamo identificare X con un punto della sfera di raggio unitario e dimensio-

ne 2k-1. Questa sfera è denominata spazio delle pre-fome, ed è indicata da S_k^3;

6. La sfera delle pre-forme viene poi identificata con lo spazio delle matrici su cui il gruppo speciale ortogonale SO (k) agisce da sinistra. Definiamo quindi Σ_k^3, cioè lo spazio delle forme di k punti in 3 dimensioni, come lo spazio quoziente della sfera delle pre-forme quozientato per il gruppo speciale ortogonale:

$$\Sigma_k^3 = S_k^3 / SO(k).$$

In questo modo anche l'effetto delle rotazioni viene filtrato, e la definizione dello spazio delle forme ottenuta.

Dopo aver definito lo spazio delle forme, ad ogni categoria sarà associata la sua forma, sulla base dei dati disponibili per ogni periodo archeologico sulla 'forma' del ritrovamento. Si noti che le forme possono essere specificate con qualche incertezza: questo è molto utile nella pratica archeologica, perché la forma dei ritrovamenti non sempre è nota con precisione. Matematicamente parlando, è possibile tenere conto dell'incertezza definendo una distanza sullo spazio delle forme, in modo tale da poter formalizzare il concetto di forma 'approssimativa': fissato un punto nello spazio delle forme, nei punti vicini ad esso sono rappresentate le forme che sono approssimativamente quella che abbiamo fissato. Qui il termine 'vicino' è da intendere sulla base della distanza definita sullo spazio delle forme.

4. Conclusioni e sviluppi futuri

In questo report abbiamo discusso alcuni problemi implementativi, relativi all'applicazione dell'algoritmo di page rank per la determinazione del potenziale archeologico. Il report è una continuazione del primo (Bini, 2011), dove gli autori hanno proposto l'applicazione di algoritmi di tipo page rank, che sono in gran parte utilizzati per classificare e dare 'importanza' alle pagine web in base ai link che inviano e ricevono. L'utilizzo del page rank è motivato dal fatto che, astraendo, le relazioni tra i ritrovamenti sono l'elemento più importante che contribuisce al potenziale archeologico, e che queste relazioni mostrano proprietà simili ai link negli algoritmi di page rank.

Naturalmente, devono essere apportate delle modifiche, in modo da adattare l'algoritmo di page rank al calcolo del potenziale archeologico. Questo report riguarda le principali modifiche:

• la determinazione delle categorie di ritrovamenti, che è necessaria per un'efficiente memorizzazione dei dati archeologici e per un efficace calcolo algoritmico del potenziale;

• il modo di costruire la matrice dei pesi nell'algoritmo di page rank, per il quale abbiamo proposto di utilizzare le relazioni spaziali e funzionali fra (categorie di) ritrovamenti.

Per quanto riguarda le proprietà spaziali abbiamo proposto di memorizzare la forma di ciascuna categoria di ritrovamenti, intesa in senso matematico, i.e. la descrizione geometrica della parte dello spazio occupato, astraendo dalla posizione, dall'orientamento e dalle dimensioni. In seguito ogni particolare realizzazione di una categoria può essere specificata attraverso alcuni parametri che descrivono la posizione, l'orientamento e le dimensioni. Per quanto riguarda le relazioni funzionali, abbiamo proposto di costruire una matrice i cui elementi indicano la probabilità della presenza di una categoria, data la presenza di un'altra categoria.

Per quanto riguarda i futuri sviluppi, un passo importante sarà rappresentato dal modo di immagazzinare e utilizzare tutte le informazioni - la forma e la dislocazione di ciascuna categoria, le relazioni funzionali tra le categorie, le informazioni geologiche - per produrre una sola matrice dei pesi, e per implementare l'algoritmo di page rank. Molte prove dovranno essere fatte per ottimizzare la scelta tra le diverse possibilità per gli algoritmi di page rank, e per assegnare così i valori più adatti per i pesi.

Bibliografia

Anichini F. 2005, *Tutela, ricerca, valorizzazione del patrimonio archeologico: progetto per il G.I.S. della città di Pisa*, t.d.l. Università di Pisa.

Anichini F., Fabiani F., Gattiglia G., Gualandi M.L. 2012, Un database per la registrazione e l'analisi dei dati archeologici, in MapPapers 1-II, pp.1-20.

Bini D., Dubbini N., Steffè S. 2011, *Modelli matematici per la determinazione del potenziale archeologico, in MapPapers 4-I, pp.68-76 .*

Broadbent S.R. 1980, *Simulating the ley-hunter (with discussion)*, Journal of Royal Statistical Society Ser. A, 143, pp. 109-140.

Kendall D.J. 1989, *A survey of the statistical theory of shape,* Statistical Science 4(2), pp. 87-120.

Kendall D.J., Barden D., Carne T.K., Le H. 1999, *Shape and shape theory, Wiley series in probability and statistics*.

Langville A.N., Meyer C.D. 2006, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press.