1 Formula base

Essendo:

- a_k il $k$esimo articolo,
- u_i l'$i$esimo utente,
- A_i le bookmarks per l'utente u_i: $A_i = \{a_k / a_k \text{ è taggato da } u_i\}$,
- $r_{a_k u_i}$ il valore di rank di a_k per u_i,
- $P_{\mathrm{H}}(U, A, B)$ la probabilità $P(x' \geq x)$ con $x = \sharp A \cap B$ considerando come indipendente la probabilità che un elemento a dell'universo U appartenga ad A or B (secondo un modello ipergeometrico o un'approssimazione binomiale).

propongo la seguente formula per calcolare $r_{a_k u_i}$:

$$r_{a_k u_i} = -\sum_{u_j \neq i} \log\left(P_{\mathrm{H}}\left(A_i \cap A_j\right)\right) \delta_{a_k \in A_j} \tag{1}$$

Il significato pratico di questa formula è il seguente: ogni utente u_j che ha nel suo set A_j (bookmark) l'articolo a_k contribuisce al rank di a_k per u_i in modo proporzionale (al log del) all'inverso della probabilità che le bookmarks di u_j siano indipendenti da quelle di u_i.

1.1 Multiple testing

Espresso in questo modo ogni utente u_j con $A_j \cap A_i \neq \emptyset$ porta dell'informazione. Tuttavia, essendo il numero di utenti elevato e pari al numero di test statistici effettuati (a fissato u_i), può accadere che la probabilità $P_{\mathrm{H}}\left(A_i \cap A_j\right)$ sia piccola per caso. Il modo più semplice per abbattere questo rumore che entrerebbe nella sommatoria è applicare la ricetta di Bonferroni che prevede di mandare a 1 la probabilità quando essa supera una certa soglia pari a $N\lambda$ con λ pari alla significatività desiderata del test (es 0.5) e N pari al numero di utenti. Siccome è noto che questa correzione è eccessivamente conservativa quello che si potrebbe più proficuamente fare è stimare l'N empirico che ottimizza i risultati.

2 Formula ricorsiva

L'effetto della formula proposta può essere interpretato nel seguente modo: inizialmente l'informazione che si ha su ciascun a_k a fissato u_i è la presenza o meno di a_k in A_i quindi si può immaginare associato a ciascun a_k un $r_{a_k u_i} = 1$ se $a_k \in A_i$, 0 altrimenti. Calcolando $r_{a_k u_i}$ si passa da un sistema a 2 stati ad un sistema contino: ogni a_k, dentro e fuori A_i, ha possibilmente un peso diverso.

Supponiamo di iterare il processo, mantenendo fissi gli A_j; sia $r_{a_k u_i}(t)$ il rank al tempo t:

$$r_{a_k u_i}(t) = -\sum_{u_j \neq i} r_{a_k u_j}(t-1) \log\left(P_{\mathrm{H}}\left(A_i \cap A_j\right)\right) \delta_{a_k \in A_j} \tag{2}$$

essendo $r_{a_k u_j}(0) = 1$ se $a_k \in A_j$, 0 altrimenti.

Nel primo passo passiamo da un sistema a 2 stati ad un sistema continuo, nei passi successivi raffiniamo via via il rank, usando ad ogni step il rank stimato al passo precedente.

In formalismo matriciale si può scrivere

$$\vec{r}_{u_i}(t) = M\vec{r}_{u_i}(t-1) \tag{3}$$

con $M_{\alpha\beta} = -\log\left(P_{\mathrm{H}}\left(A_\alpha \cap A_\beta\right)\right)$. Essendo $P_{\mathrm{H}}\left(A_\alpha \cap A_\beta\right) = P_{\mathrm{H}}\left(A_\beta \cap A_\alpha\right)$ la matrice è simmetrica, inoltre si ha $M_{\alpha\beta} \geq 0 \; \forall \; \alpha, \beta$, quindi esiste un set completo di autovettori. Sia λ l'autovalore maggiore, cerchiamo $\vec{r}_{u_i}$ tale che:

$$\vec{r}_{u_i} = \lambda M\vec{r}_{u_i}$$

L'esistenza degli autovettori significa anche che il sistema dinamico discreto $\vec{r}_{u_i}(t) = \lambda M\vec{r}_{u_i}(t-1)$ descritto sopra ha punti stazionari. Vedi wikipedia alla voce Arnoldi iteration e Lanczos algorithm per vedere come il processo iterativo converge agli autovettori. In analogia con PageRank, il vettore $\vec{r}_{u_i}$ ricercato è quello associato all'autovalore massimo (che equivale forse a quello più stabile).

3 Ricorsione con matrice variabile nel tempo

3.1 Criteri per la variazione delle bookmarks

Si possono introdurre dei criteri di variazione delle bookmarks $A_i(t)$ sulla base di $\vec{r}_{u_i}(t-1)$.

Criterio del massimo e minimo. Dato il complementare delle bookmarks $\bar{A}_i = U - A_i$ supponiamo che al tempo t si abbia:

$$\max_{\bar{A}_i}(r_{a_k u_i}) > \min_{A_i}(r_{a_k u_i})$$

allora potremo decidere di scambiare l'articolo $a_k \in \bar{A}_i$ che genera il massimo rank in $\bar{A}_i$ con quello $a_l \in A_i$ che genera il minimo rank in A_i.

Scala tipica in $\vec{r}$. Se da una analisi teoria o empirica risultasse che i valori delle componenti del vettore $\vec{r}_i$ non si distribuiscono in modo graduale tra il massimo e il minimo ma si formano due set distinti di componenti, qualcuna con un valore nettamente maggiore di un certo valore di taglio c e altre nettamente al di sotto (con poche componenti con valore attorno a c); allora si potrebbe usare c come discriminante e assumere che gli articoli in $A_i(t)$ sono quelli con $r_{a_k u_i}(t-1) > c$.

Qualunque sia il criterio con cui A_i e quindi M varia nel tempo, il problema diventa

$$r_{a_k u_i}(t) = -\sum_{u_{j\neq i}} r_{a_k u_j}(t-1) \log\left(P_{\mathrm{H}}\left(A_i(t-1) \cap A_j(t-1)\right)\right) \delta_{a_k \in A_j(t-1)}$$

Forse in matematica questo problema è noto e sarebbe interessante vedere se (date le proprietà di $M(t)$ e della funzione $f(\vec{r}_i, M) \mapsto M'$ con cui M evolve)

potessimo affermare che esiste una M di equilibrio e vedere quanto il vettore di rank all'equilibrio differirebbe dal caso in cui M è tenuta costante.

In questa formulazione l'obbiettivo non sarebbe tanto quello di calcolare il vettore di rank all'equilibrio quanto ricavare direttamente dalla dinamica gli A_i di equilibrio.

4 Disambiguare sui tag.

Prendiamo due utenti u_i e u_j supponiamo che ciascuno abbia nelle proprie bookmarks articoli che si riferiscono a due diversi ambiti di interesse. Assumiamo che gli ambiti di interesse siano caratterizzati da due tag diversi da ciascu untente. Tuttavia non facciamo affidamento sulla morfologia del tag, ovvero assumiamo che ogni utente possa utilizzare termini diversi per indicare lo stesso ambito di interesse. In tutto abbiamo 4 tag diversi (se fossero uguali morfologicamente considereremmo il fatto come accidentale e assumeremmo comunque i tag come diversi) associati a 4 ambiti di interesse; questi ultimi però possono non essere diversi, anzi assumiamo che uno dei 2 ambiti di interesse di u_i coincida con uno dei 2 ambiti di interesse di u_j. Definiamo $A_{i\tau}$ come $A_i \supseteq A_{i\tau} = \{a_k / a_k$ è taggato da u_i con il tag $\tau\}$ e assumiamo che $A_{i\tau}$ contenga articoli tutti afferenti allo stesso ambito di interesse, eventualmente parzialmente (ma non molto) sovrapposto ad altri ambiti di interesse caratterizzati da altri tag. In questo caso è probabile che il numero di elementi in ciascuna intersezione $A_{i\alpha} \cap A_{i\beta}$ sia diverso: ci aspettiamo una maggiore intersezione per quella coppia di tag $\alpha \in T_{u_i}$ e $\beta \in T_{u_j}$ che caratterizzano lo stesso ambito di interesse (dove abbiamo indicato con T_{u_i} l'insieme dei tag utilizzati da u_i). Analogamente, sulla base delle size relative dei vari $A_{i\tau}$, $P_H(U, A_i \cap A_j)$ potrebbe essere non significativa e contemporaneamente potrebbe esistere una coppia di tag $\alpha \in T_{u_i}$ e $\beta \in T_{u_j}$ tali che $P_H(U, A_{i\alpha} \cap A_{j\beta})$ sia significativa.

Proponiamo quindi un nuovo modo di calcolare il rank che tenga conto di queste considerazioni sui diversi ambiti di interesse $A_{i\tau}$da cui A_i può essere composta:

$$r_{a_k i} = - \sum_{j, \alpha \in T_{u_i}, \beta \in T_{u_j}} \log \left(P_{\mathrm{H}} \left(A_{i\alpha} \cap A_{j\beta} \right) \right) \delta_{a_k \in A_{j\beta}}$$

Abbiamo qui esteso, considerando i tag, solo la formula di base, ma analoghe estensioni si possono fare per la formula ricorsiva e quella con matrice variabile nel tempo.