Esercizio 2-layers Belief Network

Dati i seguenti documenti:

D1: **Extended boolean model** is an **information retrieval** method which has several shortcomings, e.g., there is no inherent notion of **document ranking**, and it is very hard for a user to form a good **boolean query** request.

D2: In the **vector space method** or **vector space model** text is represented by a vector of *terms*. The **cross-product** (or **dot-product**) between two vectors is often used as a **similarity measure** and is used for **relevance ranking**.

D3: This family of **information retrieval** models is based on the general principle that documents in a collection should be ranked by decreasing probability of their relevance to a query. This is often called *the **probabilistic ranking** principle*

D4: In this model, **document retrieval** and **query processing** are modeled as an **inference process** in an **inference network**.

E dato il vocabolario:

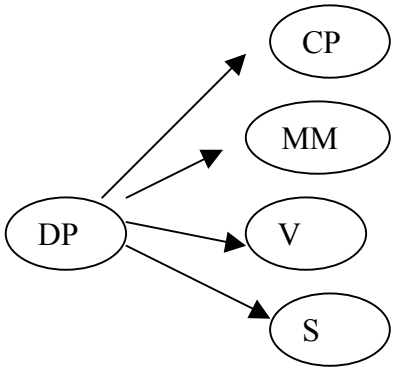
boolean query, cross product, document ranking, dot product, extended boolean model, inference network, inference process, information retrieval, inner product, probabilistic ranking, query processing relevance ranking, similarity measure, vector space method , vector space model

Costruite la rete delle dipendenze "two layers" per la query:
relevance ranking, cross product, probabilistic retrieval, query processing.

Le dipendenze sono ipotizzate da Google Sets, i cui risultati sono riportati nel seguito.
Assumete, considerando l'output di Google Sets, che il primo termine rappresenti un termine **condizionante** la presenza di tutti quelli che seguono, ed assumere p(ti/tj)=0,2 per tutti termini condizionati: ad esempio, p(cross product/dot product)=0.2 , perchè da Google Sets assumiamo che, dato "dot product" i termini ad esso correlato includano "cross product"(si veda la rete di dipendenze per dot product).

Stimate la p(Di/q) e ordinate per rilevanza.

Predicted Items
dot product
cross product
matrix multiplication
vectors
subtraction



Predicted Items
information retrieval
xml
data mining
performance
algorithms
clustering
indexing
multimedia
query processing

Predicted Items
vector space model
generalized vector space model
latent semantic indexing aka latent semantic analysis
extended boolean model

Predicted Items
probabilistic retrieval
physicalmind institute
pilates method exercise
method inc.
boundary element method
social filtering
vector space method
relevance ranking

Soluzione. La rete è la seguente (ai termini vengono sostituite le iniziali, IR=information retrieval ecc.):

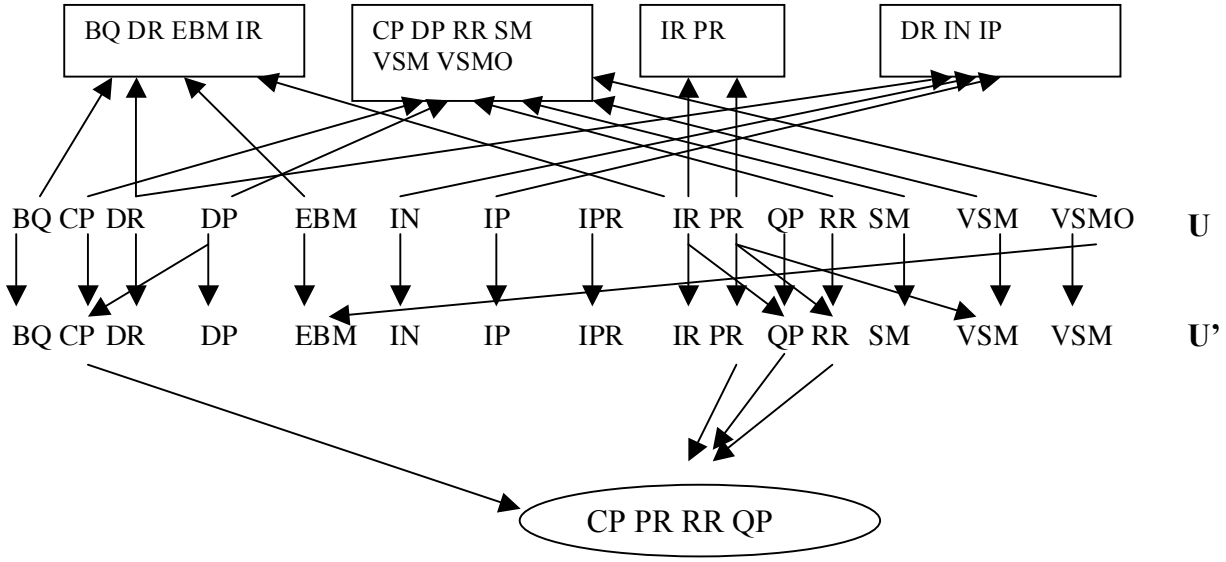


Fig. 1

Le formule per il modello Belief Networks "Two Layers" sono:

$$P(u)=\left(\frac{1}{2}\right)^{t} \qquad P(d_j \mid u)=\frac{\sum_{i=1}^{t} w_{ij}\times w_{iu}}{\sqrt{\sum_{i=1}^{t} w_{ij}^{\,2}}\times\sqrt{\sum_{i=1}^{t} w_{iu}^{\,2}}} \qquad P(q \mid u)=\sum_{\forall u'} P(q \mid u')\times P(u' \mid u)$$

$$P(d_j \mid q)=\eta\sum_{\forall u} P(d_j \mid u)\left(\sum_{\forall u'} P(q \mid u')\times P(u' \mid u)\right)\times P(u) \qquad P(q \mid u')=\begin{cases}1 & \text{if } \forall k_i, g_i(q)=g_i(u')\\ 0 & \text{otherwise}\end{cases}$$

$$P(u' \mid u)=\frac{1}{m}\sum_{\substack{\forall k_i' \in u' \\ pa(k_i')\in u}} P(k_i' \mid pa(k_i'))$$

dove m=|u'|

poichè P(u) costante per ogni u, possiamo assimilarlo in η

Inoltre, associamo ad u un vettore $\vec{k}=w1\vec{k}_1+w2\vec{k}_2+....+wt\vec{k}_t$ dove i k_i sono vettori unitari (l'argomento i-simo di $\vec{k}_i$ è 1 e tutti gli altri sono zero). Infine, supponiamo che i pesi w_i siano binari e che $p(u)=\prod_i w_i k_i$. Analogamente per u', cioè: le dipendenze condizionali esistono fra il livello u e il livello u', ma NON fra keywords di uno stesso livello.

D1

$$p(D1/u)=\frac{1}{\sqrt{4}\sqrt{6}},\quad p(D2/u)=\frac{3}{\sqrt{6}\sqrt{6}},\quad p(D3/u)=\frac{2}{\sqrt{2}\sqrt{6}},\quad p(D4/u)=0$$

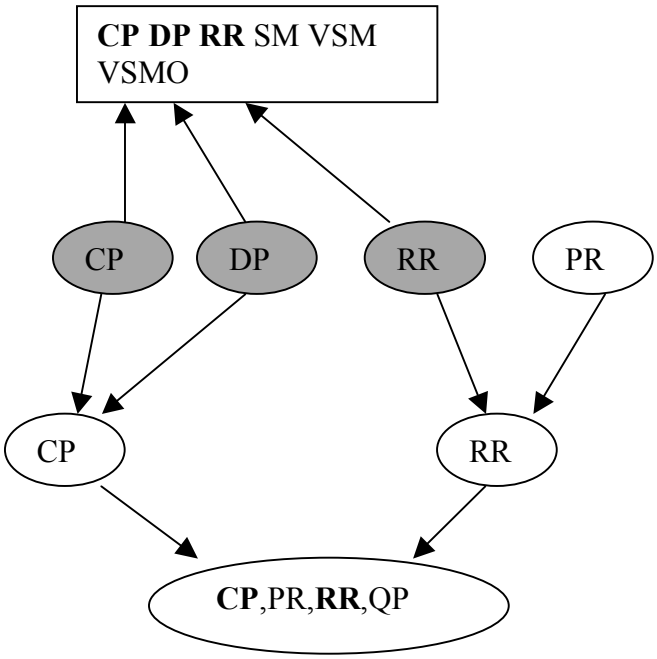
$$u: CP, DP, IR, PR, QP, RR$$

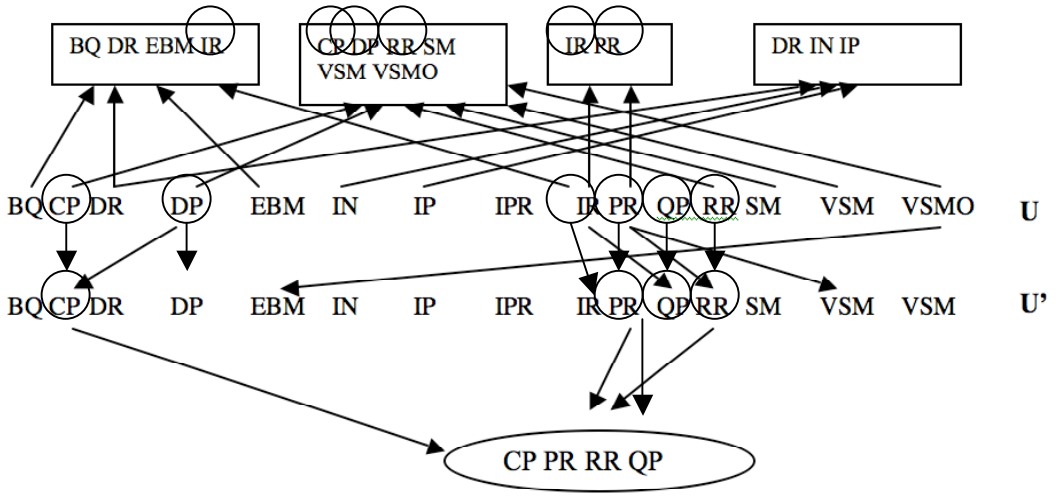
$$D1(u'=QP): P(u'/u)=\frac{1}{1}P(QP/IR)=0,2$$

$$D2(u'=CP,RR): P(u'/u)=\frac{1}{2}[(P(CP/CP,DP)+(P(RR/RR)]=\frac{1}{2}(1+1)=1$$

$$D3(u'=QP,PR,RR)=\frac{1}{3}(1+0,2+0,2)$$

Esempio della sottorete “attiva” per il calcolo del rank di D2 (in U i nodi grigi sono i soli per cui i relativi ki sono diversi da zero):





u'=CP PR RR QP (**cross product, probabilistic retrieval, relevance ranking, query processing**).

La figura mostra che la catena di dipendenze "attiva" nel set **u** i concetti: CP,DP,IR,PR,QP,RR.
Poichè $p(u) = \prod_{k_i \in u} p(k_i)$ (in **u** i concetti sono statisticamente indipendenti)

abbiamo: $p(D_1/u) = \frac{1}{\sqrt{4}\sqrt{6}}$ $p(D_2/u) = \frac{3}{\sqrt{6}\sqrt{6}}$ $p(D_3/u) = \frac{2}{\sqrt{2}\sqrt{6}}$ $p(D_4/u) = \frac{0}{\sqrt{3}\sqrt{6}}$

$p(k_i/k_i) = 1$ $p(CP/DP) = 0,2$ $p(QP/IR) = 0,2$ $p(RR/PR) = 0,2$
inoltre: $p(CP) = CP,DP$ $p(PR) = PR$ $p(QP) = QP,IR$ $p(RR) = PR,RR$