Recuperação de Informação B

Modern Information Retrieval

Cap. 2: Modeling

Section 2.8 : Alternative Probabilistic Models

September 20, 1999
Alternative Probabilistic Models

- Probability Theory
  - Semantically clear
  - Computationally clumsy

- Why Bayesian Networks?
  - Clear formalism to combine evidences
  - Modularize the world (dependencies)

Bayesian Network Models for IR
- Inference Network (Turtle & Croft, 1991)
- Belief Network (Ribeiro-Neto & Muntz, 1996)
Bayesian Inference

Schools of thought in probability

- freqüentist
- epistemological
Bayesian Inference

Basic Axioms:

- $0 \leq P(A) \leq 1$;
- $P(\text{sure}) = 1$;
- $P(A \lor B) = P(A) + P(B)$ if $A$ and $B$ are mutually exclusive.
Bayesian Inference

Other formulations

- \( P(A) = P(A \land B) + P(A \land \neg B) \)
- \( P(A) = \sum_{\forall i} P(A \land B_i) \), where \( B_{i,\forall i} \) is a set of exhaustive and mutually exclusive events
- \( P(A) + P(\neg A) = 1 \)
- \( P(A|K) \) belief in \( A \) given the knowledge \( K \)
- if \( P(A|B) = P(A) \), we say: \( A \) and \( B \) are *independent*
- if \( P(A/B \land C) = P(A/C) \), we say: \( A \) and \( B \) are conditionally independent, given \( C \)
- \( P(A \land B) = P(A|B)P(B) \)
- \( P(A) = \sum_{\forall i} P(A \mid B_i)P(B_i) \)
Bayesian Inference

**Bayes’ Rule** : the heart of Bayesian techniques

\[ P(H|e) = \frac{P(e|H)P(H)}{P(e)} \]

Where,
- \( H \) : a hypothesis and \( e \) is an evidence
- \( P(H) \) : prior probability
- \( P(H|e) \) : posterior probability
- \( P(e|H) \) : probability of \( e \) if \( H \) is true
- \( P(e) \) : a normalizing constant, then we write:

\[ P(H|e) \sim P(e|H)P(H) \]
Bayesian Networks

**Definition:**
Bayesian networks are directed acyclic graphs (DAGs) in which the nodes represent random variables, the arcs portray causal relationships between these variables, and the strengths of these causal influences are expressed by conditional probabilities.
Bayesian Networks

$y_i$: parent nodes (in this case, root nodes)
$x$: child node
$y_i$ cause $x$
$Y$ the set of parents of $x$
The influence of $Y$ on $x$
can be quantified by any function $F(x,Y)$ such that
$\sum_{x} F(x,Y) = 1$
$0 \leq F(x,Y) \leq 1$
For example, $F(x,Y) = P(x|Y)$
Bayesian Networks

Given the dependencies declared in a Bayesian Network, the expression for the joint probability can be computed as a product of local conditional probabilities, for example,

\[ P(x_1, x_2, x_3, x_4, x_5) = P(x_1) \cdot P(x_2|x_1) \cdot P(x_3|x_1) \cdot P(x_4|x_2, x_3) \cdot P(x_5|x_3). \]

\( P(x_1) \): prior probability of the root node
Bayesian Networks

In a Bayesian network each variable $x$ is conditionally independent of all its non-descendants, given its parents.

For example:

$$P(x_4, x_5 | x_2, x_3) = P(x_4 | x_2, x_3) P(x_5 | x_3)$$
Inference Network Model

- Epistemological view of the IR problem
- Random variables associated with documents, index terms and queries
- A random variable associated with a document $d_j$ represents the event of observing that document
Inference Network Model

**Nodes**
- documents ($d_j$)
- index terms ($k_i$)
- queries ($q, q_1, \text{ and } q_2$)
- user information need ($I$)

**Edges**
- from $d_j$ to its index term nodes $k_i$ indicate that the observation of $d_j$ increase the belief in the variables $k_i$.
Inference Network Model

d_j has index terms k_2, k_i, and k_t
q has index terms k_1, k_2, and k_i
q_1 and q_2 model boolean formulation
q_1 = ((k_1 \land k_2) \lor k_i);
I = (q \lor q_1)
Inference Network Model

Definitions:
- $k_1, d_j,$ and $q$ random variables.
- $k = (k_1, k_2, ..., k_t)$ a $t$-dimensional vector
- $k_i, \forall i \in \{0, 1\}$, then $k$ has $2^t$ possible states
- $d_j, \forall j \in \{0, 1\}; \forall q \in \{0, 1\}$

The rank of a document $d_j$ is computed as $P(q \land d_j)$
- $q$ and $d_j$ are short representations for $q=1$ and $d_j=1$
- ($d_j$ stands for a state where $d_j = 1$ and $\forall i \neq j \Rightarrow d_i = 0$, because we observe one document at a time)
Inference Network Model

\[ P(q \wedge d_j) = \sum_{\forall k} P(q \wedge d_j \mid k) P(k) \]
\[ = \sum_{\forall k} P(q \wedge d_j \wedge k) \]
\[ = \sum_{\forall k} P(q \mid d_j \wedge k) P(d_j \wedge k) \]
\[ = \sum_{\forall k} P(q \mid k) P(k \mid d_j) P(d_j) \]

\[ P(\neg(q \wedge d_j)) = 1 - P(q \wedge d_j) \]
Inference Network Model

As the instantiation of $d_j$ makes all index term nodes mutually independent $P(k \mid d_j)$ can be a product, then

$$P(q \land d_j) = \sum_{\forall k} \left[ P(q \mid k) \right.$$ 

$$\left. \left( \prod_{\forall i \mid g_i(k) = 1} P(k_i \mid d_j) \right) \left( \prod_{\forall i \mid g_i(k) = 0} P(-k_i \mid d_j) \right) \right] P(d_j) \right]$$

remember that: $g_i(k) = 1$ if $k_i = 1$ in the vector $k$ and $0$ otherwise
The prior probability $P(d_j)$ reflects the probability associated to the event of observing a given document $d_j$

- Uniformly for $N$ documents
  $$P(d_j) = \frac{1}{N}$$
  $$P(\neg d_j) = 1 - \frac{1}{N}$$

- Based on norm of the vector $d_j$
  $$P(d_j) = \frac{1}{|d_j|}$$
  $$P(\neg d_j) = 1 - \frac{1}{|d_j|}$$
Inference Network Model

For the Boolean Model

\[
P(d_j) = \frac{1}{N}
\]

\[
P(k_i \mid d_j) = \begin{cases} 1 & \text{if } g_i(d_j) = 1 \\ 0 & \text{otherwise} \end{cases}
\]

\[
P(\neg k_i \mid d_j) = 1 - P(k_i \mid d_j)
\]

\[\Rightarrow\] only nodes associated with the index terms of the document \(d_j\) are activated
Inference Network Model

For the Boolean Model

\[
P(q \mid k) = \begin{cases} 
1 & \text{if } \exists q_{cc} \mid (q_{cc} \in q_{dnf}) \land (\forall k, g_i(k) = g_i(q_{cc})) \\
0 & \text{otherwise}
\end{cases}
\]

\[
P(\neg q \mid k) = 1 - P(q \mid k)
\]

⇒ one of the conjunctive components of the query must be matched by the active index terms in \( k \)
Inference Network Model

For a tf-idf ranking strategy

\[ P(d_j) = \frac{1}{|d_j|} \]

\[ P(\neg d_j) = 1 - \frac{1}{|d_j|} \]

⇒ prior probability reflects the importance of document normalization
Inference Network Model

For a tf-idf ranking strategy

\[ P(k_i \mid d_j) = f_{i,j} \]

\[ P(\neg k_i \mid d_j) = 1 - f_{i,j} \]

⇒ the relevance of the a index term \( k_i \) is determined by its normalized **term-frequency** factor \( f_{i,j} = \frac{\text{freq}_{i,j}}{\text{max freq}_{i,j}} \)
Inference Network Model

For a tf-idf ranking strategy

Define a vector \( k_i \) given by

\[
\begin{align*}
  k_i &= k \mid ((g_i(k)=1) \land (\forall j \neq i \ g_j(k)=0)) \\
  \implies & \quad \text{in the state } k_i \text{ only the node } k_i \text{ is active and all the others are inactive}
\end{align*}
\]
Inference Network Model

For a tf-idf ranking strategy

\[
P(q | k) = \begin{cases} 
  \text{idf}_i & \text{if } k = k_i \land g_i(q) = 1 \\
  0 & \text{if } k \neq k_i \lor g_i(q) = 0 
\end{cases}
\]

\[
P(\neg q | k) = 1 - P(q | k)
\]

\(\Rightarrow\) we can sum up the individual contributions of each index term by its normalized \text{idf}
For a tf-idf ranking strategy

As \( P(q|k)=0 \quad \forall k \neq k_i \), we can rewrite \( P(q \land d_j) \) as

\[
P(q \land d_j) = \sum_{\forall k_i} \left[ P(q|k_i) P(k_i|d_j) \right. \\
\left. \prod_{l \neq i} P(\neg k_l|d_j) \right] P(d_j) \\
= \left( \prod_{\forall k_i} P(\neg k_i|d_j) \right) P(d_j) \\
\sum_{\forall k_i} \left[ P(k_i|d_j) P(q|k_i) / P(\neg k_i|d_j) \right]
\]
For a tf-idf ranking strategy

Applying the previous probabilities we have

\[ P(q \land d_j) = C_j \frac{1}{|d_j|} \sum_{i} [f_{i,j} \cdot \text{idf}_i \left( \frac{1}{1 - f_{i,j}} \right)] \]

\[ \Rightarrow C_j \text{ vary from document to document} \]

\[ \Rightarrow \text{the ranking is distinct of the one provided by the vector model} \]
Inference Network Model

Combining evidential source

Let \( I = q \lor q_1 \)

\[
P(I \land d_j) = \sum_{k} P(I \mid k) P(k \mid d_j) P(d_j)
\]

\[
= \sum_{k} [1 - P(\neg q \mid k) P(\neg q_1 \mid k)] P(k \mid d_j) P(d_j)
\]

⇒ it might yield a retrieval performance which surpasses the retrieval performance of the query nodes in isolation (Turtle & Croft)
Belief Network Model

- As the Inference Network Model
  - Epistemological view of the IR problem
  - Random variables associated with documents, index terms and queries

- Contrary to the Inference Network Model
  - Clearly defined sample space
  - Set-theoretic view
  - Different network topology
Belief Network Model

The Probability Space

Define:

\(K=\{k_1, k_2, ..., k_t\}\) the sample space (a concept space)
\(u \subset K\) a subset of \(K\) (a concept)
\(k_i\) an index term (an elementary concept)

\(k=(k_1, k_2, ..., k_t)\) a vector associated to each \(u\) such that
\(g_i(k)=1 \iff k_i \in u\)

\(k_i\) a binary random variable associated with the index term \(k_i\), \((k_i = 1 \iff g_i(k)=1 \iff k_i \in u)\)
Belief Network Model

A Set-Theoretic View

Define:

- a document $d_j$ and query $q$ as concepts in $K$
- a generic concept $c$ in $K$
- a probability distribution $P$ over $K$, as
  \[ P(c) = \sum_u P(c|u) P(u) \]
  \[ P(u) = (1/2)^t \]

$P(c)$ is the degree of coverage of the space $K$ by $c$. 
Belief Network Model

Network topology

query side

document side
Belief Network Model

Assumption

$P(d_j|q)$ is adopted as the rank of the document $d_j$ with respect to the query $q$. It reflects the degree of coverage provided to the concept $d_j$ by the concept $q$. 
Belief Network Model

The rank of $d_j$

\[ P(d_j|q) = \frac{P(d_j \land q)}{P(q)} \]

\[ \sim P(d_j \land q) \]

\[ \sim \sum_u P(d_j \land q | u) \ P(u) \]

\[ \sim \sum_u P(d_j | u) \ P(q | u) \ P(u) \]

\[ \sim \sum_k P(d_j | k) \ P(q | k) \ P(k) \]
Belief Network Model

For the vector model

Define

Define a vector $k_i$ given by

$$k_i = k | ((g_i(k)=1) \land (\forall j \neq i \ g_j(k)=0))$$

$\Rightarrow$ in the state $k_i$ only the node $k_i$ is active and all the others are inactive
Belief Network Model

For the vector model

Define

\[
P(q \mid k) = \begin{cases} 
\frac{(w_{i,q}}{|q|)} & \text{if } k = k_i \land g_i(q) = 1 \\
0 & \text{if } k \neq k_i \lor g_i(q) = 0 
\end{cases}
\]

\[
P(\neg q \mid k) = 1 - P(q \mid k)
\]

\[
\Rightarrow (w_{i,q} \mid |q|) \text{ is a normalized version of weight of the index term } k_i \text{ in the query } q
\]
Belief Network Model

For the vector model

Define

\[
P(d_j \mid k) = \begin{cases} 
(w_{i,j} / |d_j|) & \text{if } k = k_i \land g_i(d_j) = 1 \\
0 & \text{if } k \neq k_i \lor g_i(d_j) = 0
\end{cases}
\]

\[
P(\neg d_j \mid k) = 1 - P(d_j \mid k)
\]

\[\Rightarrow (w_{i,j} / |d_j|) \text{ is a normalized version of the weight of the index term } k_i \text{ in the document } d_j\]
Bayesian Network Models

Comparison

- Inference Network Model is the first and well known
- Belief Network adopts a set-theoretic view
- Belief Network adopts a clearly define sample space
- Belief Network provides a separation between query and document portions
- Belief Network is able to reproduce any ranking produced by the Inference Network while the converse is not true (for example: the ranking of the standard vector model)
Bayesian Network Models

Computational costs

- Inference Network Model one document node at a time then is linear on number of documents
- Belief Network only the states that activate each query term are considered
- The networks do not impose additional costs because the networks do not include cycles.
Bayesian Network Models

Impact

The major strength is net combination of distinct evidential sources to support the rank of a given document.