

БАЙЕСОВСКИЕ МЕТОДЫ В ЗАДАЧЕ ОЦЕНКИ РЕЛЕВАНТНОСТИ ПРИ ПОИСКЕ РАБОТЫ В INTERNET

Д.Б. Шехтер, А.В. Чадюк, А.Л. Червинский-Ивашура

ООО «Эр-Джи-Дейта Украина»
03056, Киев, Политехническая, 33, корп. 2, оф. 605
тел.: +38(044)241-91-31, факс.: +38(044)236-31-88
shekter@rgdata.com.ua; anato lyc@rgdata.com.ua; achervinsky@rgdata.com.ua;

В данной работе рассматривается применение байесовских методов для определения релевантности объявлений о работе при поиске работы через Интернет. Детально описывается байесовская сеть доверия как инструмент определения меры релевантности объявления о работе пользовательскому запросу. Формулируется постановка задачи оценки релевантности для данной предметной области, приводятся необходимые теоретические сведения о выбранной методике решения задачи. Указаны основные аспекты практической реализации системы – факторы, влияющие на релевантность объявления о работе и их учет в байесовской сети.

In this paper we deal about Bayesian methods for relevance estimation task in Internet job searching. We consider a Bayesian belief network for relevance measure calculation, and describe this network in detail. The paper defines a problem of relevance estimation in the domain, and depicts an essential theoretical background for this problem solving. This work also points to the main aspects of system implementation, such as factors that should be taken into account when Bayesian network for job leads relevance estimation is used.

Введение

Оценка релевантности – важнейшая задача любого информационного поиска. Релевантность как соответствие между поисковым запросом и найденной информацией является одним из фундаментальных понятий при поиске информации. Разработка любого поискового двигателя в сети Интернет также влечет за собой необходимость решения задачи оценки релевантности.

Рост объема информации в сети Интернет в последнее время существенно затрудняет процесс обнаружения релевантных документов и фильтрацию нерелевантных документов [1]. Особенно это касается поиска в определенной предметной области. Поисковые двигатели общего назначения, во-первых, не предоставляют пользователю возможности сужения области поиска, а во-вторых, в силу своей ориентации на работу со слабо структурированными документами (например, HTML-страницами) не в состоянии производить оценку релевантности с учетом специфики конкретной области.

Необходимость решения этих проблем привела к созданию поисковых двигателей узкого назначения – *вертикальных поисковых двигателей*. Эти поисковые системы ограничиваются поиском в определенной предметной области (вертикали), пытаются улучшить качество предоставляемых пользователю результатов. Улучшение происходит за счет поиска информации, недоступной поисковым двигателям общего назначения, и предоставления дополнительной функциональности, характерной для данной предметной области [2]. Вертикальные поисковые двигатели имеют различную архитектуру, однако для всех таких систем задача определения релевантности остается важнейшей.

В рамках данной работы рассматриваем оценку релевантности при поиске работы в сети Интернет. Рынок труда является динамичной, постоянно меняющейся предметной областью, а поиск вакансий на рынке труда – задачей информационного поиска, решение которой все большее количество людей склонно находить в сети Интернет. Построение поискового двигателя для поиска работы в сети Интернет требует решения задачи оценки релевантности, и в нашем случае эта задача была решена с помощью байесовских методов вероятностного рассуждения.

Постановка задачи

Пусть имеем некоторое множество документов (объявлений о приеме на работу), полученное из сети Интернет. Каждый документ характеризуется определенными реквизитами, типичными для объявления о работе: должность, краткое описание, требуемые навыки, текст объявления, размер заработной платы, необходимый уровень образования и т.п. Пользователь поискового двигателя указал в своем запросе ключевые слова, характеризующие подходящую для него работу. Кроме того, пользователь имеет также возможность указать определенное (возможно, пустое) множество критериев, например, месторасположение работы и желаемый размер заработной платы. Совокупность ключевых слов и критериев будем называть *запросом пользователя*.

Степень соответствия каждого конкретного документа запросу пользователя называем *релевантностью* документа запроса. Задача поискового двигателя – предоставление пользователю максимально релевантных результатов, объявлений, максимально соответствующих его запросу.

Таким образом, необходимо построить интеллектуальную систему, позволяющую определить меру релевантности каждого имеющегося у нас документа введенному запросу пользователя. *Мерой релевантности* мы называем число, в соответствии со значением которого мы можем проводить сортировку документов по релевантности. Это число должно обладать следующими свойствами:

- 1) мера релевантности - неотрицательное вещественное число;
- 2) мера релевантности тем выше, чем выше релевантность объявления запроса;
- 3) мера релевантности должна быть ограничена сверху.

Последнее условие крайне важно с точки зрения удобства пользователя. Предоставляя пользователю возможность анализировать меру релевантности как число из некоторого диапазона, мы в некоторой степени даем ему оценить абсолютную степень соответствия документа его запросу.

Поставим цель получить продемонстрировать пользователю релевантность в виде целого числа от 0 % до 100 %. Для этого мы округлим и нормируем нашу меру релевантности, подразумевая при этом, что 100 %-релевантный документ представляет собой объявление, наверняка соответствующее запросу пользователя с точки зрения нашей системы.

Байесовская сеть доверия для оценки релевантности документов

В нашем случае был реализован собственный подход к определению релевантности – интеллектуальная полнотекстовая пост-обработка найденных документов с помощью байесовской сети доверия (Bayesian belief network).

Теоретические сведения. Байесовские сети используются для моделирования предметных областей, которые характеризуются неопределенностью. Эта неопределенность может быть обусловлена недостаточным пониманием предметной области, неполным знанием ее состояния в момент принятия решения, случайным характером механизмов, определяющих поведение этой области, или комбинацией этих факторов.

Байесовские сети также называют байесовскими сетями доверия (БСД) или просто сетями доверия. Раньше использовался термин «причинные вероятностные сети». БСД – это граф, вершины которого соединены направленными ребрами, с сопоставленной каждому узлу вероятностной функцией. Сеть в БСД представляет собой направленный ациклический граф (DAG, directed acyclic graph), т.е. граф, в котором не существует направленного маршрута, начинающегося и заканчивающегося в одной и той же вершине.

Вершина БСД представляет собой или дискретную случайную величину с конечным количеством состояний, или непрерывную гауссовскую величину. В этой работе рассматривается БСД с дискретными переменными, для которых как синонимы используются слова «узел» и «вершина». Ребра между вершинами представляют причинно-следственные связи между ними.

Если вершина не имеет родителей (не существует ребер, направленных к ней), она будет содержать таблицу безусловных вероятностей своих состояний. В случае дискретной вершины такая таблица содержит распределение вероятностей между всеми возможными состояниями этой вершины. Если же у вершины есть родители (одно или несколько ребер, направленных к ней), то такая вершина содержит таблицу условных вероятностей (СРТ, conditional probability table), каждая ячейка которой содержит условную вероятность пребывания вершины в определенном состоянии для случая определенной конфигурации состояний всех ее родителей. Таким образом, количество ячеек в таблице условных вероятностей дискретной вершины БСД равно произведению количества возможных состояний этой вершины на произведение количества возможных состояний всех ее родительских вершин.

Тривиальная БСД, показана на рис. 1, отображает причинно-следственную взаимосвязь между двумя элементами некоторой предметной области – *A* и *B*. Наличие причинно-следственной связи от *A* к *B* означает нашу мысль о том, что если *A* находится в некотором состоянии, то это влияет на состояние *B*.



Рис.1. Тривиальная байесовская сеть

Случайная дискретная переменная, которую представляет вершина *A*, может находиться в одном из двух состояний – a_1 или a_2 . Вершина *B* имеет три возможных состояния: b_1, b_2, b_3 . Таблицы условных вероятностей для этих вершин имеют такой вид.

Пример таблиц условных вероятностей для вершин БСД

A	$P(a_i)$
a_1	0.5
a_2	0.5

Таблица 1

B	$P(b_i a_1)$	$P(b_i a_2)$
b_1	1	0.6
b_2	0	0.2
b_3	0	0.2

Поскольку вершина A не имеет родителей, значения вероятностей ее состояний не являются зависимыми (про распределение вероятностей вершины A в таком случае свидетельствуют, что заданы *априорные* вероятности ее состояний). Для вершины B , наоборот, вероятности состояний зависят от состояния ее родительской вершины A :

- если A находится в состоянии a_1 , то B находится в состоянии b_1 ;
- если A находится в состоянии a_2 , то вероятность того, что B находится в состоянии b_1 равна 0.6, а в состояниях b_2 и $b_3 - 0,2$.

Если вершина A не имеет родителей, то вместо условных вероятностей (автоматически) используются безусловные вероятности $P(A)$.

Именно процесс исчисления вероятностей является основой для принятия решений в условиях неопределенности на основе байесовских сетей [3]. Раскрытие неопределенности (dealing with uncertainty) осуществляется в БСД путем вычисления вероятностей состояний интересующих нас вершин на основе имеющейся информации о значении (части) других вершин сети. Математический базис для этого процесса определяет байесовский подход к анализу неопределенности и соответствующий ему аппарат классической теории вероятностей.

Основу байесовского подхода составляет понятие *условной вероятности* $P(A|B) = x$, которая означает, что при условии возникновения B (и всего остального, что не имеет отношения к B) вероятность возникновения A равняется x . Совместная вероятность наступления A и B определяется формулой полной вероятности

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A). \quad (1)$$

Уравнение (1) - *фундаментальный принцип* исчисления вероятностей и основа для *теоремы Байеса*:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \quad (2)$$

Теорема Байеса применяется, если в нашем распоряжении есть информация о зависимых переменных, а суть исследования состоит в определении вероятности исходных переменных [4]. Так, пусть известна условная вероятность $P(B|A)$ возникновения некоторого события B при условии, что имеет место событие A . Тогда теорема Байеса дает решение обратной задачи – какова вероятность возникновения события A , произошло событие B .

Действительно, пусть A_1, A_2, \dots, A_n – полная группа несовместных взаимоисключающих событий (или альтернативных гипотез). Тогда апостериорная вероятность $P(A_j|B)$ каждого из событий $A_j, j=1..n$ при условии, что произошло событие B , выражается априорной вероятностью A_j :

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{j=1}^n P(B|A_j)P(A_j)}. \quad (3)$$

Оценка релевантности документов. Одна из задач, в которой успешно применяются байесовские сети - задача классификации. Так называемый наивно-байесовский классификатор, представляющий собой простую байесовскую сеть, является одним из самых эффективных классификаторов [5]. Наш подход предполагает, что задачу оценки релевантности также можно рассматривать как задачу классификации. Действительно, рассмотрим каждый документ (объявление о работе) как принадлежащий к одной из двух непересекающихся областей: C_1 - релевантные документы, C_2 - нерелевантные документы.

В таком случае, задача оценки релевантности документа запроса представляется в виде задачи отнесения его к одному из двух классов. В этом случае, принадлежность документа к первому классу позволяет свидетельствовать, что этот документ является релевантным запроса.

Решим эту задачу с помощью байесовской сети, сопоставив понятию «документ» вершину сети. Эта вершина может находиться в двух состояниях: c_1 – «документ релевантен» и c_2 – «документ не релевантен». Априорные вероятности этих состояний положим равными 0.5, что соответствует понятию неопределенности в вероятностном анализе: $P(c_1) = P(c_2) = 0.5$. Если в результате вычислений получим, что вероятность нахождения этого узла в состоянии «документ релевантен» равна 0.9, то это будет означать, что с вероятностью 0.9 данный документ принадлежит к классу C_1 .

Далее, пусть $F = \{F_i\}, i=1..n$ – множество факторов, влияющих на релевантность документа. Рассмотрим, например, такой фактор, как наличие ключевого слова запроса в заголовке документа. Очевидно, что наличие ключевого слова в заголовке повышает релевантность документа. Тогда введем в сеть вершину F_1 , соответствующую событию «ключевое слово в заголовке документа». Эта вершина будет иметь два состояния: f_{11} – «ключевое слово встретилось в заголовке документа» и f_{12} – «ключевое слово не встретилось в заголовке

документа». Если мы знаем условные вероятности $P(f_{1j} | c_i)$, $i, j=1..2$, то у нас есть таблица условных вероятностей для вершины F_j , и мы можем рассчитать вероятности $P(c_i | f_{1j})$, $i, j=1..2$.

Для отнесения документа D к классу релевантных в случае, когда нам известно состояние f_{1j} , используется очевидное правило: если $P(c_1 | f_{1j}) > P(c_2 | f_{1j})$, то $D \in C_1$.

Следовательно, для определения релевантности мы должны выделить все факторы, составляющие множество F , и задать таблицы условных вероятностей для каждого фактора. Каждый из факторов рассчитывается соответствующим образом для каждого ключевого слова запроса.

Таким образом, вершины сети в нашем случае – это факторы, влияющие на вероятность нашего «главного» узла, отвечающего за релевантность документа в целом. Байесовская сеть для нашей задачи имеет вид.

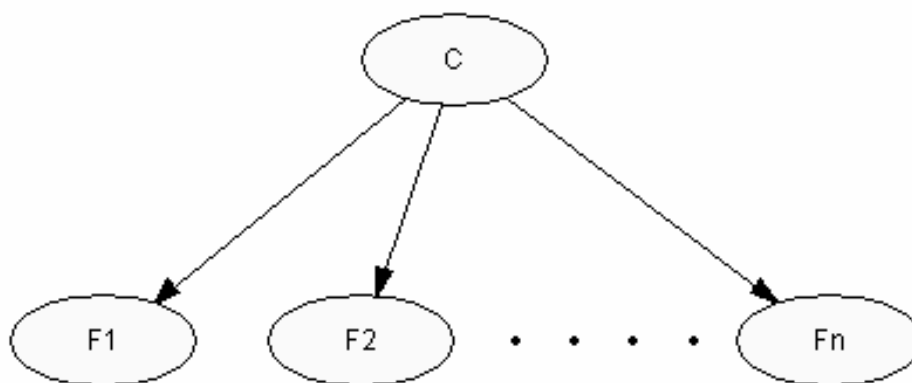


Рис. 2. Байесовская сеть для оценки релевантности документа запросу

Здесь C – вершина сети, представляющая собой вероятность того, что документ релевантен запросу, а $F_1, F_2 \dots F_n$ – факторы, учитываемые при расчете этой вероятности. Существенным моментом является направление причинно-следственных связей в сети. Так, стрелки выходят из вершины C и входят в вершины F_i . Здесь байесовская сеть выполняет обратный логический вывод – определяет вероятность каждого состояния вершины C при известных состояниях вершин F_i .

Вышерассмотренный случай фактора F_i принимал одно из двух значений. Однако факторы могут иметь различную природу – они могут принимать несколько значений, а могут вообще не быть дискретными. В общем случае рассматриваем определенный диапазон изменения значений каждого фактора. Пусть некоторый фактор F_i принимает значение $x \in [x_{\min}; x_{\max}]$. Тогда мы нормируем значение этого фактора в диапазон $[-1; 1]$ с помощью формулы

$$\tilde{x} = \frac{x - \frac{x_{\min} + x_{\max}}{2}}{x_{\max} - \frac{x_{\min} + x_{\max}}{2}} \quad (4)$$

и принимаем расчетные вероятности для соответствующей вершины равными

$$\begin{aligned} \tilde{P}(f_i | c_1) &= \frac{1 - [1 - 2 \cdot P(f_i | c_1)] \cdot \tilde{x}}{2}, \quad i=1..n \\ \tilde{P}(f_i | c_2) &= \frac{1 - [1 - 2 \cdot P(f_i | c_2)] \cdot \tilde{x}}{2}, \quad i=1..n, \end{aligned} \quad (5)$$

где $P(f_i | c_1)$ – элемент таблицы условных вероятностей для i -ой вершины сети, показывающий, с какой вероятностью в релевантном документе фактор F_i принимает максимальное значение $x = x_{\max}$; $P(f_i | c_2)$ – вероятность, с которой фактор F_i принимает максимальное значение $x = x_{\max}$ в нерелевантном документе.

Полученные расчетные вероятности $\tilde{P}(f_i | c_1)$ и $\tilde{P}(f_i | c_2)$ теперь можно использовать в формуле Байеса:

$$P(c_1 | f_i) = \frac{P(f_i | c_1) \cdot P(c_1)}{P(f_i | c_1) \cdot P(c_1) + P(f_i | c_2) \cdot P(c_2)}, \quad i=1..n. \quad (6)$$

Отметим, что $\tilde{P}(f_i | c_1) = P(f_i | c_1)$ при $x = x_{\max}$, и $\tilde{P}(f_i | c_1) = 1 - P(f_i | c_1)$ при $x = x_{\min}$. Для остальных значений $x \in (x_{\min}; x_{\max})$ расчетная вероятность $\tilde{P}(f_i | c_1) \in (1 - P(f_i | c_1); P(f_i | c_1))$. Это означает, что возрастание значения фактора x приводит к последовательному (линейному) возрастанию значения соответствующей расчетной вероятности.

Таким образом, вышеописанная схема позволяет учесть как дискретные, так и непрерывные значения факторов, влияющих на общую релевантность документа. При этом, если возрастание значения фактора соответствует *уменьшению* релевантности (например, количество дней, прошедших с даты публикации объявления), то достаточно повторить эти рассуждения для случая, когда элемент таблицы условных вероятностей $P(f_i | c_1)$ показывает, с какой вероятностью в релевантном документе фактор F_i принимает *минимальное* значение $x = x_{\min}$.

В нашем случае сеть является достаточно тривиальной, чтобы расчет вероятностей мог быть выполнен последовательным применением теоремы Байеса. Разумеется, такой расчет возможен только в том случае, если мы делаем сильное предположение об *условной независимости* вершин сети. Условная независимость вершин байесовской сети означает блокирование влияния между этими вершинами. Переменные (множества переменных) F_1 и F_2 являются независимыми при известном состоянии переменной A , если

$$P(F_1 | A) = P(F_1 | A, F_2). \quad (7)$$

Это означает, что если состояние вершины A известно, то никакая информация о F_1 не изменяет вероятности F_2 . В случае нашей сети это представляется отсутствием причинно-следственных связей между всеми факторами множества F .

На самом деле это предположение, очевидно, является абсолютно нереалистичным (именно поэтому классификаторы подобной структуры и носят название «наивных»). В то же время нарушение этого предположения в условиях реального мира не обнаруживает существенного влияния на конечный результат. Оказывается, что такой последовательный подход является в нашем случае преимуществом, так как резко снижает вычислительную сложность и соответственно скорость работы алгоритма.

Свидетельствуя о вопросе получения численных значений для таблиц условных вероятностей, следует отметить, что концептуально для решения этой задачи выделяются два подхода [6]:

- получение информации от экспертов предметной области;
- получение информации на основании данных.

Таблицы условных вероятностей чаще всего генерируются на основании данных с помощью статистических методов. Однако стоит отметить, что принципиально субъективный байесовский подход не требует «объективности» вероятностей, а поэтому позволяет при формировании таблиц условных вероятностей опираться на субъективные оценки экспертов. Условные вероятности, численные значения которые мы используем для расчета, получены на основании слияния результатов статистических исследований и экспертных оценок. Мы провели статистический анализ множества релевантных и нерелевантных документов для разных запросов по разным источникам информации и занесли эти значения в таблицы условных вероятностей сети.

О практической реализации

При реализации системы мы выделили такие факторы, влияющие на релевантность объявления о приеме на работу:

Приведенные в табл. 2 факторы мы представили в виде вершин байесовской сети, каждая из которых может принимать соответствующие состояния, и задали таблицы условных вероятностей для этих вершин. При поступлении поискового запроса система выполняет расчет каждого фактора для каждого ключевого слова и выполняет распространение соответствующих расчетных вероятностей в сети. Результатом работы является вероятность $P(C | F_1, F_2 \dots F_n)$ для каждого имеющегося документа D , которая и является мерой релевантности документа запроса.

Если $P(C = c_1 | F_1, F_2 \dots F_n) > P(C = c_2 | F_1, F_2 \dots F_n)$, то документ релевантен запросу, т.е.

$$P(C = c_1 | F_1, F_2 \dots F_n) > 0.5, \text{ то } D \in C_1. \quad (8)$$

Документы, удовлетворяющие решающему правилу (8), выводятся пользователю с нормированной мерой релевантности

$$P = \frac{P(C | F_1, F_2 \dots F_n) - P_{\min}}{P_{\max} - P_{\min}} \cdot 100\% = 2 \cdot [P(C | F_1, F_2 \dots F_n) - 0.5] \cdot 100\%. \quad (9)$$

Факторы, введенные в байесовскую сеть в качестве вершин

Таблица 2

Фактор	Различаемые состояния	Пояснение
Вхождение ключевого слова в заголовок документа	1) 0 вхождений 2) 1 и более вхождений	Наличие ключевого слова в <i>position title</i> повышает релевантность объявления; отсутствие понижает релевантность объявления
Вхождение ключевого слова в краткое описание работы	1) 0 вхождений 2) Ровно 1 вхождение	Наличие ключевого слова в <i>summary</i> (первые 25 слов объявления) повышает релевантность объявления; отсутствие не изменяет релевантности объявления
Многочисленные вхождения ключевого слова в краткое описание работы	1) Менее 2 вхождений 2) 2 и более вхождений	Вхождение ключевого слова в <i>summary</i> два и более раз повышает релевантность объявления
Количество вхождений ключевого слова в текст объявления о работе	1) Менее 2 вхождений 2) От 2 до 7 вхождений 3) Более 7 вхождений	Наличие в тексте объявления ключевого слова 2 и более раз повышает релевантность объявления (нелинейно, по дискретным значениям фактора «2», «3», «4», «5», «6», «7 и более»); наличие ровно одного вхождения не изменяет релевантности объявления, отсутствие вхождений понижает релевантность
Положение вхождения ключевого слова в текст документа	Значения в интервале от 0.6086 до 1	Положение слова в тексте документа представляется числом от 0 (конец документа) до 1 (начало документа); большее значение этого числа повышает релевантность объявления (нелинейно, по непрерывным значениям фактора)
Количество вхождений биграмм (пар слов) в заголовок документа	1) 0 вхождений 2) 1 и более вхождений	Наличие в <i>position title</i> фразы (биграммы), совпадающей с фразой из двух ключевых слов, повышает релевантность объявления; отсутствие фразы не изменяет релевантности объявления
Количество вхождений биграмм в полный текст объявления	1) 0 вхождений 2) От 1 до 4 вхождений 3) Более 4 вхождений	Наличие в полном тексте объявления (<i>summary + text + skills</i>) биграммы 1 и более раз повышает релевантность объявления (нелинейно, по дискретным значениям фактора «1», «2», «3», «4 и более»); отсутствие фразы не изменяет релевантности объявления
Значение фактора TF*IDF для ключевого слова	1) 0 2) Значения в интервале от 0 до 4 3) Значения более 4	Большее значение фактора TF*IDF [7], учитывающего частоту вхождения ключевого слова (TF) и вес слова в документе (IDF), повышает релевантность объявления (нелинейно, по непрерывным значениям фактора в интервале от 0 до 4, при значении «4 и более» - максимально); значение 0 не изменяет релевантности объявления
Дата публикации объявления	Значения в интервале от 0 до 50 (дней)	Фактор представляет количество дней, прошедшее с момента публикации объявления и до текущей (сегодняшней) даты. Большее значение фактора понижает релевантность объявления (нелинейно, по дискретным значениям «1», «2», ... «49», «50 и более»); значение 0 («сегодня») не изменяет релевантности объявления

Заклучение

Таким образом, суть нашего подхода к анализу релевантности объявлений о работе состоит в использовании байесовской сети доверия. Мы адаптировали механизм вероятностного принятия решений для оценки релевантности, представив эту задачу как задачу классификации документа – отнесения его к классу релевантных или нерелевантных. Такая классификация осуществлена на основании расчета вероятности принадлежности документа к той или иной категории. Эта же вероятность выступает и *мерой релевантности*, позволяя нам отсекаать объявления с низкой релевантностью, сортировать множество полученных результатов, предоставлять пользователю возможность выбора порога релевантности и так далее.

Предлагаемый подход успешно применен в рамках практической реализации поискового двигателя для поиска объявлений о работе в Интернет. Дальнейшее развитие применяемой методики может заключаться в разработке персональных агентов для поиска работы, обладающих механизмами адаптации численных значений таблиц условных вероятностей для каждого конкретного пользователя.

1. *Chau M.* Spidering and Filtering Web Pages for Vertical Search Engines / Proceedings of The Americas Conference on Information Systems.- AMCIS 2002 Doctoral Consortium, Dallas, TX, USA, 2002.
2. *Steele R.* Techniques for Specialized Search Engines // Proceedings of Internet Computing '01.- Las Vegas, NV, USA, 2001.
3. *Jensen F.V.* Bayesian networks basics // Tech. Rep. Department of Mathematics and Computer Science, Aalborg University, Denmark, 1996. – 12 p.
4. *Терехов С.А.* Введение в байесовы сети // Научная сессия МИФИ-2003. V Всероссийская научно-техническая конференция „Нейроинформатика-2003”: лекции по нейроинформатике. Часть 1.- М.:МИФИ, 2003. - С. 149-187.
5. *Friedman N., Goldszmidt M.* Building classifiers using Bayesian networks. In Proceedings of the National Conference on Artificial Intelligence '96.- AAAI Press, Menlo Park, CA, USA, 1996. - P. 1277 -1284.
6. *Kaerulff U., Jensen F.V.* Bayesian networks / Tech. Rep. Department of Computer Science, Aalborg University, Denmark, 1996. - 5 p.
7. *Salton G., McGill M.J.* An Introduction to Modern Information Retrieval.- McGraw-Hill Publishing, New York, NY, USA, 1983.