7-я лекция. Эффективность информационных систем

На прошедших лекциях мы рассматривали принципы действия информационных систем. Теперь сосредоточимся на вопросах оценки эффективности этого действия.

Функция ИПС состоит в выделении из поискового массива таких документов, которые содержат информацию, удовлетворяющую **информационную потребность** пользователя. Но информационная потребность выражается в **информационном запросе**, формулировка которого может лишь более или менее приблизительно выражать действительную информационную потребность. («Мысль изреченная есть ложь»). Информационный запрос представляется поисковой системе в виде **поискового образа запроса (ПОЗ)** – формализованного перечня терминов. Кроме того задаётся формальный **критерий соответствия (КС)** документа запросу. Поисковый образ запроса вместе с критерием соответствия составляют поисковое предписание: **ПП = ПОЗ + КС**. Поисковая система в ответ на запрос, выполняя поисковое предписание, выдаёт некоторую совокупность документов.

Однако не все документы в выдаче удовлетворяют информационную потребность. Как правило, они лишь формально соответствуют поисковому предписанию. Документы, действительно соответствующие потребности пользователя, называются **пертинентными**. А сама информационная потребность представляет собой весьма сложное психическое явление, и проблема повышения степени пертинентности выдачи оказывается не только

трудной для достижения, но её даже трудно чётко поставить как практическую задачу. Определить соответствие выдачи запросу проще. Документы, соответствующие запросу, называются **релевантными**. Однако суждение о релевантности будет зависеть от того, кто это суждение выносит. Если автор запроса, то он будет оценивать не столько релевантность, сколько пертинентность, в той мере, в которой ему удастся ознакомиться с документом. Если же релевантность будет оценивать работник системы, то он сможет объективно учитывать только формальное вхождение элементов поискового предписания в документ, не задаваясь вопросом о соответствии запроса потребности пользователя («Каков запрос, таков и ответ»). Но именно последняя характеристика определяет эффективность самой системы. Таким образом надо различать **техническую релевантность** и **семантическую релевантность** (соответствие смыслу, а не форме запроса).

Для организации выдачи документов система должна уметь оценивать релевантность априори, **до** выдачи, чтобы выдать именно релевантные документы. А для определения качества работы системы оценку релевантности выданных документов производят апостериори, **после** выдачи. Конечно апостериорная релевантность сильно зависит от априорной, но эти характеристики различны по своей природе. Так, при автоматическом поиске система не имеет ничего, кроме поискового предписания и поисковых образов документов. Это значит, что система может устанавливать только соответствие этих объектов, что влечёт введение ещё одного параметра – **системной релевантности.**

Простейший и весьма распространённый критерий релевантности состоит в требовании полного совпадения поискового образа документа с поисковым предписанием. Но этот критерий применим только к ограниченным видам запросов, например к поиску по полному библиографическому описанию, или к поиску всех документов в некотором тематическом классе по принятой классификации знаний. А в реальных поисковых системах при всестороннем координатном индексировании вероятность полного совпадения предписания с поисковым образом документа крайне низка. Поэтому необходимо как-то оценивать не абсолютную, а относительную релевантность – степень релевантности – на основе частичного совпадения поискового предписания с поисковым образом документа. При этом система должна выдавать документ, если степень его релевантности запросу превзошёл некоторый достаточно высокий порог. Методов вычисления степени системной релевантности было предложено довольно много, и многие из них имеют весьма изощрённый характер в попытке по формальным признаками промоделировать человеческое восприятие сходства и различия смысла текстов. Рассмотрим некоторые из них.

Начнём с простых оценок. Степенью релевантности можно считать отношение числа дескрипторов запроса, найденных в документе A к общему числу N дескрипторов в запросе: $R_1 = A/N$. Требование полного совпадения

запроса с документом соответствует $R_1 = 1$ и A = N = M, где M – полное число дескрипторов в поисковом образе документа. В практических поисковых системах порог релевантности задают установлением допустимой разницы (d) между общим числом дескрипторов в запросе N и числом их, найденных в документе. Если поиск на полное совпадение даёт неудовлетворительный результат, проводят поиск на совпадение всех, кроме одного дескриптора (d=1), кроме двух (d=2) и т. д. Если же запрос состоит всего из одного термина, то естественно – можно вести поиск только на полное совпадение.

Более сложный случай, когда за критерий релевантности принимается величина $R_2 = A/M$ – отношение числа найденных дескрипторов в документе к числу всех дескрипторов в поисковом образе документа. Требование полного совпадения здесь также соответствует $R_2 = 1$, но как показала практика, для систем с таким критерием релевантности удовлетворительная выдача наблюдается при установлении порога выдачи в пределах от R_2=0,25 до $R_2 = 0{,}4$. Очевидно, что R_2 зависит от принятой глубины и разносторонности индексирования документов, от среднего числа M дескрипторов в поисковом образе документа. При многословном поисковом образе документа и запрос также должен быть многословным. Если M = 10, то поиск по **одному** понятию никогда не даст $R_2 > 0{,}1$ и система ничего не выдаст. В запрос придётся добавлять новые термины, как бы объясняя системе свою потребность.

Эти два описанных критерия релевантности можно усложнить учётом значимости дескрипторов для документа и для запроса, если этим дескрипторам в процессе индексирования присвоены весовые коэффициенты. Пусть в документе совпали дескрипторы № 1, 2, 3, …, k. Пусть этим дескрипторам пользователь присвоил веса $n_1, n_2, n_3, \ldots, n_k$, а в документе они имеют веса $m_1, m_2, m_3, \ldots, m_k$. Тогда в качестве критерия релевантности можно принять сумму произведений этих весовых коэффициентов: $R_3, = m_1n_1 + m_2n_2 + m_3n_3 + \ldots + m_kn_k$, или как кратко пишут математики:

$$R_3 = \sum_{i=1}^{k} m_i n_i$$

Однако, для того, чтобы релевантность не зависела от масштабов присвоения коэффициентов, величину R_3 следует взять относительно общей суммы всех коэффициентов дескрипторов в запросе и/или в документе:

$$R_4 = \left(\sum_{i=1}^{k} m_i n_i\right) \Big/ \left(\sum_{ПОД} m_i \cdot \sum_{ПОЗ} n_i\right),$$

где в знаменателе суммы берутся по всем дескрипторам поискового образа документа (ПОД) и поискового образа запроса (ПОЗ) соответственно.

От конкретной формулы расчёта релевантности, принятой в информационной системе, эффективность поиска зависит в сильной степени.

В одной из американских информационных систем Министерства обороны ещё в 50-х годах прошлого века была реализована изощрённая процедура расчета релевантности, при которой для каждого термина запроса просматривался весь имеющийся массив документов (ПОД) и подсчитывалась частота совместной встречаемости данного термина со всеми другими. Далее для каждого термина составлялся упорядоченный список (профиль) терминов совместно встречающихся чаще, чем в среднем (связанные термины). Далее из всех профилей терминов запроса выбираются общие для всех их. С отобранными терминами процедура повторяется. На основе частоты совместной встречаемости терминов этого списка вычисляется их вес (чем больше связанность, тем выше вес). Наконец на основе этих весов рассчитывался показатель релевантности аналогичный R_3.

Подобные сложные расчёты статистики распределения терминов в документах имеют назначение как-то выявить смысловые связи слов. Однако возникает вопрос: «Зачем заставлять машину выяснять то, что человеку ясно априори?». Смысловые связи слов можно прямо заложить в машину в виде информационно-поискового тезауруса, о чём мы говорили в прошлый раз. Эта идея впервые была реализована в практической информационно-поисковой системе, видимо, в нашей стране. Это ИПС «Пусто – непусто», разработанная ВИНИТИ и внедрённая в ЦНТИ «Информэлектро». Ведущие разработчики – Э. С. Бернштейн и Д. Г. Лахути.

Такое, довольно странное название системы «Пусто–непусто» обусловлено принятым в ней критерием релевантности. Он определялся соотношением наполненности четырёх множеств:

M_1 - множество дескрипторов, совпадающих в ПОД и ПОЗ;

M_2 - множество дескрипторов ПОД, родовых для дескрипторов ПОЗ.

M_3 - множество дескрипторов ПОД, видовых для дескрипторов ПОЗ;

M_4 - множество дескрипторов ПОД, не связанных с дескрипторами ПОЗ (поискового образа запроса).

По соотношению пустоты и наполненности этих множеств можно ранжировать и выбирать конкретный критерий выдачи документов. Наиболее вероятна релевантность документа, если все его дескрипторы совпадают с запросом:

M_1 совпадающие	M_2 родовые	M_3 видовые	M_4 посторонние
+	0	0	0

Столь же вероятна релевантность, если в документе есть также видовые дескрипторы (может быть наряду с родовыми):

+	0	0	0
+	+	0	0

Эти документы составляют первый эшелон выдачи. Если же в документе есть только видовые дескрипторы, то это может значить, что в нём

идет речь только о части понятий, интересующих пользователя. Документы с заполненным только M_3

0	0	+	0

составят второй эшелон выдачи.

В том случае, когда в документе представлены обобщающие (родовые) понятия, это может означать, что речь там идёт об общих вещах, а конкретно интересующее пользователя понятие упоминается только как частность. Документы с заполненным M_2 составляют третью очередь выдачи.

+	+	0	0
0	+	0	0
0	+	+	0

Документы, содержащие посторонние дескрипторы ($M_4 \neq 0$) в той системе решено было не выдавать вовсе, хотя и они могли содержать полезную информацию.

Общая таблица эшелонов выдачи такова:

Эшелон	M_1 совпадающие	M_2 родовые	M_3 видовые	M_4 посторонние
Первый	+	0	0	0
	+	0	0	0
	+	+	0	0
Второй	0	0	+	0
Третий	+	+	0	0
	0	+	0	0
	0	+	+	0

Важно в этом примере не то, какой именно был выбран показатель соответствия, а то, что для его определения использованы **знания** логических связей понятий, заложенные в систему и представляющие там некоторую модель предметной области, в которой действует система. Наличие такой модели является необходимым условием интеллектуального подхода системы к своей задаче. На пути развития этой идеи прогнозируется дальнейший прогресс в разработке автоматизированных систем вообще, и информационных систем в частности.

До сих пор мы говорили о том, как **система** оценивает эффективность того или иного документа для пользователя, а теперь остановимся на том, как **пользователь** может оценить эффективность системы для себя. Вообще-то эффективность системы для заказчика определяется как её техническим качеством, так и экономическим – стоимостью, в обратно пропорциональной зависимости. Но мы пока будем говорить чисто о технической эффективности.

Степень технической эффективности может быть определена сравнением реальной действующей системы с идеальной моделью. Идеальная модель может быть определена (как это было сделано основоположником научно-технической информатики К. Муэрсом) так: Это система, которая из документального фонда выдаёт ровно те и все те документы, которые **бы отобрал** сам пользователь, если бы он мог внимательно прочитать каждый из них. В этом определении, казалось бы абсолютно ясном, при внимательном обсуждении оказывается не ясным главное слово: «Что значит **отобрал бы**»? Отбирают документ для того, чтобы ознакомиться с ним. Но если пользователь «их внимательно прочитал», то значит он их уже всех «отобрал». А если считают, что «отбор» имеет целью получение полезной для дела информации, то это зависит от конкретного **дела**, и заранее определено быть не может. Это сильно снижает ценность определения эффективности систем, которое как раз и нужно определять прежде «дела», когда идёт речь о приобретении, внедрении или разработке системы. «До дела» можно определить только эффективность относительно технической релевантности, а «входе дела» пользователь судит о системе по её реальной пертинентности, которая заведомо ниже.

Так или иначе, соотношение реальной выдачи $M_р$ с идеальной $M_и$ характеризуется следующими множествами документов:

A – документы, выданные системой, входящие в желаемую выдачу

$$A = M_р \cap M_и$$

B – документы, выданные системой, не входящие в желаемую выдачу

$$B = M_р \cap \neg M_и$$

C – документы, не выданные системой, но входящие в желаемую выдачу

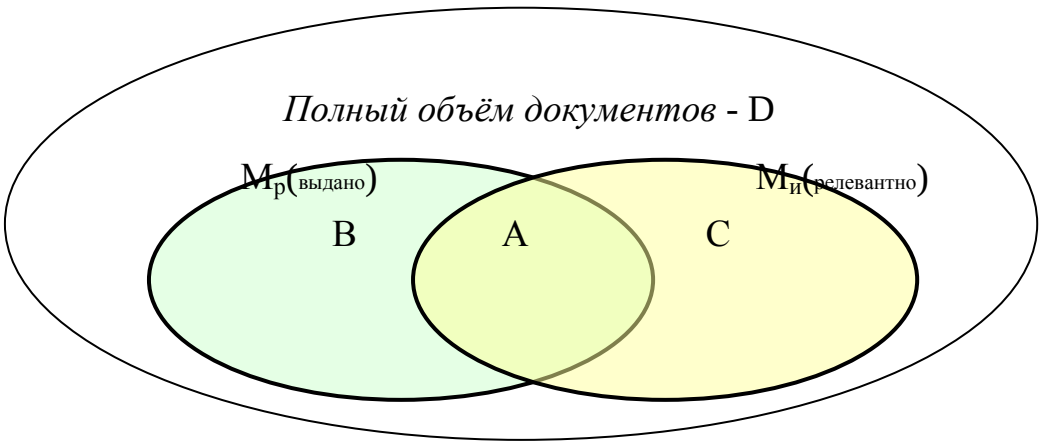
$$C = \neg M_р \cap M_и$$

D– документы, не входящие ни в реальную, ни в желаемую выдачу

$$D = \neg M_р \cap \neg M_и$$

(Знак ¬ здесь означает дополнение множества до полного объёма документов и читается как отрицание «*не*»).

В идеальном случае $M_р = M_и = A$, $B = C = Д = 0$



Реальный случай может характеризоваться соотношением числа документов в этих множествах. n_a - число документов во множестве A, n_b - число документов в B, n_c - число документов в C, n_d - число документов в D.

Наиболее популярны два отношения, это:

- **Коэффициент точности** $T = n_a /(n_a+n_b)$ - отношение числа релевантных документов в выдаче к общему объёму выдачи.

- **Коэффициент полноты** $П = n_a /(n_a+n_c)$ - отношение числа релевантных документов в выдаче к общему числу релевантных документов в массиве.

Множество документов в выдаче, не соответствующих запросу B называется **шумом** (информационный шум). Относительное количество шумовых документов в выдаче $Ш = n_b /(n_a+n_b)$ называется **коэффициентом шума**. Ш + Т = 1.

Рассматриваются и другие коэффициенты, имеющие смысл как мера вероятности выдачи документов разной степени релевантности.