УДК 681.324.06

Ю.В. Рогушина

ИСПОЛЬЗОВАНИЕ КРИТЕРИЕВ ОЦЕНКИ УДОБОЧИТАЕМОСТИ ТЕКСТА ДЛЯ ПОИСКА ИНФОРМАЦИИ, СООТВЕТСТВУЮЩЕЙ РЕАЛЬНЫМ ПОТРЕБНОСТЯМ ПОЛЬЗОВАТЕЛЯ

Рассматриваются критерии, по которым оценивается удобочитаемость текста. Предложены пути их использования для персонификации информационного поиска в Интернет.

Введение

Существуют различные алгоритмы и методы, позволяющие информационно-поиско-вым системам (ИПС) по запросу пользователя, описывающему искомый информационный ресурс (ИР), и множеству описаний ИР, известных системе, найти наиболее подходящий ИР (или набор таких ИР). При этом могут сильно различаться средства представления запросов (запрос может содержать ключевые слова, фрагменты текста, параметры искомого ИР, его тип и т.д.), описания ИР (запись в индексе поисковой системы, метаописание, полный текст документа), а также алгоритмы их сопоставления (учитывать или нет расстояние между ключевыми словами, какой вес присваивать разным элементам метаописания, обрабатывать ли наличие ссылок на другие ИР и т.п.). При поиске могут использоваться знания соответствующей предметной области (ПрО), интересующей пользователя, и эвристические правила, позволяющие учитывать семантику ИР.

Информационное пространство Web как набора данных имеет ряд принципиальных особенностей [1] по сравнению с традиционными ИПС, которые разрабатывались и тестировались на относительно небольших и однородных коллекциях:

- объем информации, представленной в Web, на много порядков превышает объем информационных ресурсов, например электронных библиотек;
- динамичность информационных ресурсов Web очень высока;
- связи между ИР не согласованы из-за отсутствия какого-либо контроля;
- качество информации, представленной в Web, никем не контролируется и поэтому информация может быть некорректной (например, уже устаревшей), ложной, плохо сформулированной, содержать ошибки (опечатки, грамматические ошибки и т.п.).

Эти особенности вызывают потребность в создании более совершенных поисковых механизмов, позволяющих не только обрабатывать формальный запрос, заданный пользователем, но и учитывать дополнительные сведения о самом пользователе и сфере его информационных потребностей.

Персонификация поиска в Интернете. В Интернете изменяется понятие "типичного пользователя": в отличие от традиционных ИПС, где пользователь вначале обучается правилам работы с системой, принципами составления запросов, языкам их представления и т.п., большинство пользователей ИПС Интернет характеризуются следующими свойствами:

- неумение правильно и четко сформулировать запрос (типичные запросы очень коротки, лишь немногие используют расширенный поиск и логические выражения; пользователи часто допускают ошибки в запросах – как грамматические, так и логические);
- разнообразие в знаниях и потребностях пользователей очень велико, поэтому практически невозможно создать универсальную ИПС, удовлетворяющую всех;

- пользователь не готов долго изучать результаты, полученные от ИПС (так, больше половины пользователей просматривают только первую страницу с результатами поиска, и, таким образом, важно, чтобы ссылки на наиболее релевантные потребностям пользователя ИР находились в первой десятке найденных).

Разнородность контингента пользователей ИПС в Интернет обуславливает разнородность их информационных потребностей [2]. Как сле0434ствие, это обуславливает существование множества модифицированных вариантов задачи поиска. Поэтому важно связать процедуру поиска с особенностями конкретного пользователя.

Традиционно считается, что цель пользователя ИПС – обнаружение документов, содержащих информацию, релевантную его запросу. Поскольку представление пользователя о том, что такое релевантный документ, напрямую зависит от цели, для достижения которой он проводит поиск, то естественной кажется идея оптимизировать метод поиска под эту конкретную цель. Типичным примером изменения цели поиска является сужение области поиска на документы определенной категории, такой, как, например, множество домашних страниц или множество анонсов научных конференций.

Более важным параметром, по которому следует оценивать эффективность ИПС Интернет, является не релевантность результатов поиска конкретному запросу, а его пертинентность – то, насколько найденные ИР способны удовлетворить реальные информационные потребности пользователя. Чтобы повысить пертинентность поиска, необходимо учитывать знания как о ПрО поиска, так и сведения о пользователе – в частности, его способность воспринимать найденную информацию.

Известно много метрик, характеризующих ИР: свежесть, доступность, авторитетность, популярность, точность, тематическая направленность и т.д. [3]. Подобные оценки используются многими глобальными ИПС для определения порядка предоставления пользователю результатов поиска (например, Google). Наиболее известным расширением индекса цитирования в Web является PageRank, который определяет важность страницы рекурсивно на основе информации о ссылающихся на нее страницах. Общим недостатком этих оценок является то, что они ориентированы не среднестатистического пользователя и не учитывают персонифицированных сведений о нем.

Важное значение имеет пространственно-временной контекст ИР: например, запрашивая прогноз погоды, пользователь обычно подразумевает сведения о погоде в том городе, где он живет, и на сегодня, а не в другом полушарии и на позавчера.

Если пользователю сложно правильно сформулировать запрос, отражающий его информационную потребность, то он может попытаться воспользоваться поиском по образцу. При этом предполагается, что у пользователя уже есть один или несколько документов-образцов, а целью поиска обычно является не обнаружение почти идентичных или синтаксически близких страниц, а обнаружение тематически близких страниц. К сожалению, большинство существующих ИПС, хотя и поддерживают такой режим поиска, но выполняют его крайне плохо (особенно – для мультимедийных объектов).

Возможность использования информации о специфике конкретного пользователя дает возможность лучше обслуживать его потребности. Персональная информация о пользователе [4] может включать как явно указанные пользователем предпочтения, так и информация, полученная на основе анализа его предыдущего поведения (запросов, просмотренных документов и т.п.). Следует отметить, что персонифицированные подходы к поиску, реализованные в различных ИПС, не учитывают то, насколько найденный ИР понятен конкретному пользователю. Это зависит как от самого ИР (насколько сложным языком изложен материал, используется ли специальная терминология и т.п.), так и от образовательного уровня самого пользователя, причем многое зависит от того, в какой сфере находятся специальные

знания пользователя: например, для врача будет непонятен специализированный текст по физике, а понятность текста на иностранном языке во многом определяется не только знаниями пользователя в соответствующей ПрО, но и уровнем владения этим языком.

Постановка задачи. Для эффективного удовлетворения информационных потребностей пользователя важно, чтобы ИПС предлагала ему в первую очередь те ИР, которые наиболее соответствуют его познаниям в определенной ПрО, – достаточно понятные, но содержащие новую информацию. Для этого необходимо учитывать специфику ПрО, формальные характеристики исследуемых ИР (такие, как уровень удобочитаемости) и персональные сведения о пользователе (например, его образовательный уровень и компетентность в соответствующей ПрО). Ряд современных интеллектуальных ИРС позволяет задавать контекст поиска и учитывать предысторию запросов пользователя, но они не позволяют пользователю явно указать "уровень сложности" искомого ИР. Предлагается учитывать при выполнении поисковой процедуры критерии удобочитаемости текста для того, чтобы найденные ИР не только были релевантны запросу, но и соответствовали уровню компетентности пользователя. Это избавит пользователя от необходимости просматривать как слишком примитивные для него ИР, так и те, которые слишком сложны для него.

Критерии удобочитаемости информации. Одним из важных факторов персонификации поиска ИР следует считать степень понятности полученной информации пользователю. Удобочитаемость (readability) – мера доступности информации, содержащейся в тексте. Она зависит от того, насколько легко понять содержание текста (используются ли в тексте сложные термины, длинные предложения и т.п.), а также от факторов, которые не связаны с содержанием текста, например, типа шрифта, его размера и цвета. Следует различать формальную удобочитаемость текста $R_{form}(I)$, являющейся функцией только от параметров самого ИР I, и его индивидуальную удобочитаемость $R_{ind}(I,u)$, которая зависит как от характеристик ИР I, так и от свойств читателя u. Удобочитаемость можно рассматривать как комбинацию характеристик текста и читательских качеств. В общем случае она зависит от характеристик текста, таких как качества печати (в электронной версии ИР – это разборчивость шрифта, выбор цвета текста и фона); наличия иллюстраций; лексики; концептуальных трудностей; синтаксиса; структурной организации ИР; и характеристик читателя, таких как способности читателя и его интерес к материалу, изложенному в ИР.

К способностям читателя относят его образовательный уровень, компетентность в определенной предметной области, владение языком, на котором изложен материал, и т.д. Например, текст, перевод которого на родной язык читателя воспринимается им достаточно легко, в оригинале может читаться со значительными усилиями и потерей значительной части смысла. Многие ИПС позволяют пользователю указывать язык искомых ИР, но не позволяют указывать уровень своей компетентности.

Интерес читателя – важный фактор, но он, к сожалению, плохо поддается оценке. Очень важна внутренняя мотивация читателя, сфера его интересов, поэтому в выборе ИР, интересных конкретному пользователю, необходимо каким-либо образом формализовать описание такой сферы. Проще определить внешнюю мотивацию пользователя, в частности, указать полезность ИР для выполнения определенных работ, прохождения тестов и изучения курсов из различных областей знаний. Например, ИР, в названии которого фигурирует слово “учебник“ или “пособие“, предпочтительнее ИР, в названии которого использовано слово “реферат“ или “аннотация“ (при прочих равных условиях).

Лексику принято считать лучшим показателем удобочитаемости текста.

Средняя длина слов (в буквах или символах) и предложений являются статистическими факторами, которые часто используют в формулах для оценки удобочитаемости. Эти параметры легко поддаются количественному выражению и пригодны для автоматической оценки. К сожалению, существующие ИПС, предоставляя пользователю различные другие формальные параметры найденных по его запросу ИР, не выводят никакой информации об этих свойствах. Представляется целесообразным использовать оценки этих свойств для персонификации работы ИПС.

Методы оценки формальной удобочитаемости информации. Существует целый ряд методов, с помощью которых можно оценить сложность написанного. Основные крите-рии, которые при этом учитываются, – общее количество слов в тексте, средняя длина предложения и длина используемых слов. Их можно использовать и для персонификации работы с пользователем ИПС.

Подобные тесты анализируют длину предложений и слов в них, но не учитывают структуру предложений и порядок слов в них. Формулы, используемые подсчета, отличаются для различных языков (например, в английском языке слова в среднем короче, чем в украинском или русском – за счет использования окончаний). Подсчет количества слогов в различных языках осуществляется по-разному и основывается на произношении слов. Например, в английском языке (в отличие от украинского и русского) количество слогов может не совпадать с количеством гласных в слове. Например, в слове “surface“ два слога, а в слове “surfaces“ – три. Аббревиатуры обычно оцениваются по правилу “количество букв совпадает с количеством слогов“, а стандартные сокращения (км, кг и т.д.) оценивают в один слог. Так как заголовки и подзаголовки обычно не являются полными предложениями, то при анализе их игнорируют. Процедура обработки формул также неоднозначна, поэтому и они обычно игнорируются [5].

Приведем несколько примеров распространенных методов определения удобочитаемости текста [6] и рассмотрим, какие параметры влияют на уровень понятности текста. Константы, используемые в формулах для оценки сложности англоязычного текста в большинстве случаев не совпадают с теми, которые применяют для оценки украино- и русскоязычного текста. Однако полезным является уже само ознакомление с принципами подхода к определению уровня читабельности текста, а в дальнейшем возможна и научная адаптация этих подходов к оценке русскоязычных текстов специалистами).

Индекс туманности Ганнинга. Один из наиболее популярных методов оценки удобочитаемости текстовой информации – "индекс туманности" ("fog index"), разработанный в 1952 году американским ученым Р. Ганнингом [7]. Он позволяет определить минимальный возраст читателя, которому будет понятен данный текст. Используется этот индекс для оценки текстов, ориентированных на широкую аудиторию, и предполагает некоторые среднестатистические оценки образовательного уровня и интеллекта читателей. Индекс туманности измеряет сложность чтения, исходя из средней длины предложения и процента слов, состоящих из трех и более слогов.

Чем выше индекс туманности, тем сложнее читать текст. Для оценки выбирается как минимум два произвольных фрагмента текста, содержащие приблизительно по 100 слов. Учитывается средняя длина предложения (в словах) и среднее количество слогов в словах. Параметры, по которым определяется индекс туманности:

- общее количество слов в тексте k;
- количество предложений в тексте s;
- средняя длина предложения w;
- среднее количество "длинных" слов (более трех слогов) l.

Сначала определяют w – среднее количество слов в одном предложении количество слов $w_i, i = \overline{1,s}$ в предложениях делится на количество полных предложений s, $w = \sum_{i=1}^{s} \frac{w_i}{s}$, где w_i – количество слов в i-м предложении. Затем подсчитывается l

– среднее количество слов, имеющих три и больше слогов $l = \sum_{i=1}^{s} \frac{l_i}{s}$, где l_i – количество "длинных" слов в i-м предложении. При этом простые предлоги, слова, написанные прописными буквами (аббревиатуры) и слова во множественном числе или являющиеся производными, не учитываются. Подсчет индекса туманности определяется по формуле

$$F_{Gunning} = \sum_{i=1}^{s} \frac{w_i}{s} + 0.4 \sum_{i=1}^{s} \frac{l_i}{w_i} = w + 0.4l . \quad (1)$$

Этой формулой пользуются, чтобы понять, будет ли понятен текст определенной аудитории. Например, для текстов, понятных большинству населения, индекс туманности должен быть ниже 12 (для текстов на английском языке).

Формула Флеща. Формула определения легкости чтения английского текста, разработанная Р. Флещем [8], позволяет установить уровень удобочитаемости текста и приблизительный уровень образования, необходимый для того, чтобы понять написанное. Она основана на подсчете не только числа слов в предложении, но и числа слогов в каждом слове. Тест FRES (Flesch Reading Ease Score) получил широкое распространение после принятия в ряде штатов США законодательных норм, требующих, чтобы текст договора страхования мог быть понятен лицам со средним образованием. Обрабатывается фрагмент текста размером приблизительно в 100 слов. При этом учитываются такие параметры:

- общее количество слов в тексте k;
- количество предложений в тексте s;
- общее количество слогов в тексте f;
- средняя длина предложения w;
- средняя длина слова p (в слогах).

Аббревиатуры, символы и слова, написанные через дефис, рассматриваются как отдельные слова. Уровень удобочитаемости оценивается по формуле

$$F_{Flesch} = 206{,}835 - (\frac{k}{s} * 1{,}015 + \frac{f}{k} * 84{,}6) =$$
$$= 206{,}835 - (w * 1{,}015 + p * 84{,}6). \quad (2)$$

«Понятный английский язык» имеет индекс F_{Flesch} не ниже 60, а разговорный английский язык – примерно 80. Текст с индексом $F_{Flesch} \geq 90$ понятен школьникам 4–5 класса, а при $F_{Flesch} \leq 30$ – сложен для восприятия даже людям с высшим образованием.

Формула Флеща–Кинкэйда. Эта формула преобразует оценку (2) в уровень образования, необходимый для понимания оцениваемого текста. Она используется преподавателями, библиотекарями и т.д. для выбора рекомендуемых книг и учебников:

$$F_{Flesch-Kincaid\_grade} = \\ = w * 0{,}39 + p * 11{,}8 - 15{,}59. \quad (3)$$

$$F_{Flesch-Kincaid\_age} = w * 0{,}39 + p * 11{,}8 - 10{,}59 . \quad (4)$$

При этом учитываются такие параметры:

- средняя длина предложения w;
- средняя длина слова p.

График читабельности текста по Фраю. Обрабатывается, как и в предыдущих оценках, фрагмент текста размером приблизительно в 100 слов. Строится график зависимости количества предложений и слогов в обрабатываемом фрагменте текста (рис.1). Кривая на диаграмме отображает нормальный текст.

Чтобы оценить текст в соответствии с этим графиком, необходимо учитывать [9]:

- общее количество слов в тексте k;
- количество предложений в тексте s;
- общее количество слогов в тексте f;
- средняя длина предложения w;
- средняя длина слова p (в слогах).

Точка на графике, соответствующая значениям w и p, определяет уровень читабельности текста по Фраю.

Индекс Колемана–Лиау. Индекс предназначен для оценки удобочитаемости текста. В отличие от большинства подобных оценок (кроме ARI), он базируется не на среднем количестве слогов в слове, а на среднем количестве символов в слове [10]. В нем учитываются такие параметры:

- общее количество символов в тексте x;

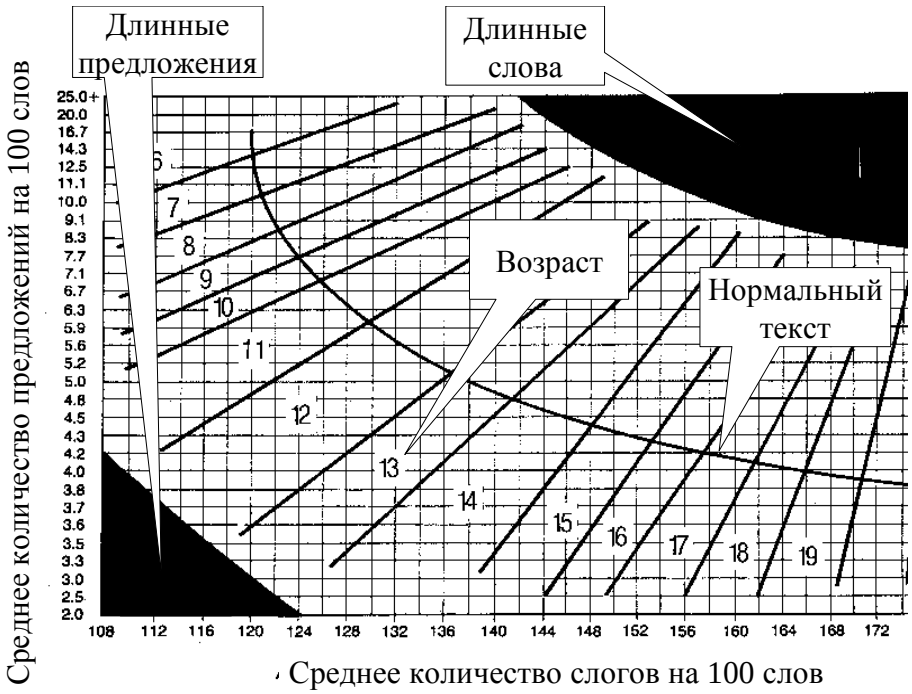

Рис. 1. График читабельности Фрая

- общее количество слов в тексте k;
- количество предложений в тексте s;
- средняя длина предложения w;
- средняя длина слова P (в символах).

Индекс Колемана–Лиау высчитывается по формуле

$$F_{\text{Coleman-Liau}} = 5{,}89 * \frac{x}{k} + 30 * \frac{s}{k} - 15{,}8. \quad (5)$$

Оценка читабельности Рэйгора. График читабельности Рэйгора [11] подобен графику оценки читабельности Фрая. В нем параметрами оценки являются средняя длина предложения и среднее количество символов в словах. Для оцениваемого текста высчитываются эти характеристики, и на пересечении соответствующих линий на графике обнаруживается возраст читателя, которому будет понятен данный текст (для текстов на английском языке). При оценке учитываются такие параметры:

- общее количество слогов в тексте f;
- общее количество слов в тексте k;
- количество предложений в тексте s;
- средняя длина предложения w;
- средняя длина слова p (в слогах).

Формула Пауэрса – Самнера – Кеарла. Данная формула используется только для текстов, предназначенных для детей [12]. Также как и в большинстве других подобных оценок, она анализирует текст размером приблизительно в 100 слов и позволяет определить возраст и образовательный уровень читателя, которому будет понятен предложенный текст:

$$F_{\text{Powers-Sumner-Kearl\_grade}} = w * 0{,}0778 + f * 0{,}0455 - 2{,}2029. \quad (6)$$

$$F_{\text{Powers-Sumner-Kearl\_age}} = w * 0{,}0778 + f * 0{,}0455 - 2{,}7971. \quad (7)$$

В оценках (6) и (7) учитываются такие параметры:

- общее количество слогов в тексте f;
- общее количество слов в тексте k;
- количество предложений в тексте s;
- средняя длина предложения w;
- средняя длина слова p (в слогах).

Формула Маклаулина "SMOG". По этой формуле анализируется фрагмент текста из 30 предложений [13], $s = 30$. Учитывается количество слов из трех и более слогов.

$$F_{SMOG\_grade} = \sqrt{L/s}\sqrt{l} + 3. \quad (8)$$

$$F_{SMOG\_age} = \sqrt{L/s}\sqrt{l} + 8. \quad (9)$$

Формулы (8) и (9) обычно дают более высокие оценки по сравнению с аналогичными тестами, потом что тест Маклаулина предназначен для прогнозирования

100% (полного) понимания текста. В оценках (8)– (9) учитываются такие параметры:

- общее количество слов в тексте k;
- количество "длинных" слов в тексте L;
- среднее количество "длинных" слов в тексте l.

Формула FORCAST. Эта формула была разработана для оценки понятности технической документации в армии США [14], поэтому она ориентирована только на взрослых читателей. Анализируемый текст может не содержать законченные предложения. В оценках (10) и (11) учитываются такие параметры:

- общее количество слов в тексте k;
- количество односложных слов в тексте b.

$$F_{FORCAST\_grade} = 20 - k * 0{,}0667 * b. \quad (10)$$

$$F_{FORCAST\_age} = 25 - k * 0{,}0667 * b. \quad (11)$$

Сравнение результатов тестов. Вышеприведенные примеры показывают разнообразие подходов к оценке уровня удобочитаемости текстов. Несмотря на то, что в формулах (1)–(11) использовались различные параметры (средняя длина слова измерялась в слогах и в символах, подсчитывалось количество многосложных и односложных слов) и коэффициенты для вычисления возраста и образовательного уровня читателя, которому будет понятен анализируемый текст, получаемые по ним оценки в целом похожи, а различия обычно вызваны особенностями авторского стиля.

Исследования показывают, что большинство читателей отдает предпочтение текстам несколько выше своего уровня. Тесты позволяют определить возрастную границу понимания, т.е. средний читатель такого возраста "на пределе" понимает изложенную в тексте информацию. Большинство оценок основывается на том, что текст считается понятным читателю, который в состоянии правильно ответить на 50 % вопросов по тексту, т.е. понимает половину информации в тексте. Следует отметить, что все эти тесты предназначены для оценки уже написанного текста, а не для рекомендаций по их написанию. Кроме того, они не учитывают порядок слов в предложении, от которого тоже в значительной степени зависит понятность текста.

Кроме того, все эти тесты дают среднестатистические оценки и не учитывают, что один и тот же человек может иметь различный образовательный уровень в разных сферах (даже при обучении в средней школе по стандартной программе ученики могут иметь различные оценки по тем или иным предметам). Поэтому при использовании таких оценок для поиска информации в Интернете представляется целесообразным использовать различные оценки в различных ПрО для одного и того же пользователя.

Использование оценки удобочитаемости информации при поиске в Интернет. Удобочитаемость текста полезно оценивать не "в общем", а с точки зрения конкретного читателя: помимо характеристик самого текста, должны учитываться характеристики самого читателя: его опыт, навыки и мотивации. Например, если читатель ищет справочную литературу по своей профессиональной деятельности, то она, вероятно, по сложности должна соответствовать профессиональному уровню читателя или даже быть выше, в то время как при необходимости получить справку по какой-нибудь непрофильной для читателя области его больше заинтересует более простой документ.

С одной стороны, грамотному читателю неинтересен слишком примитивный текст. С другой стороны, неподготовленный читатель (например, ребенок) не захочет изучать текст, который он не может воспринимать свободно (рис. 2). При этом одна и та же информация может быть изложена более сложным или простым языком. Сравним два определения: «Онтология – формальная спецификация разделяемой концептуализации, которая имеет место в некотором контексте предметной области. Она представляет собой структуру иерархически или полииерархически связанных семантических категорий, допускающая самоописание использованных в

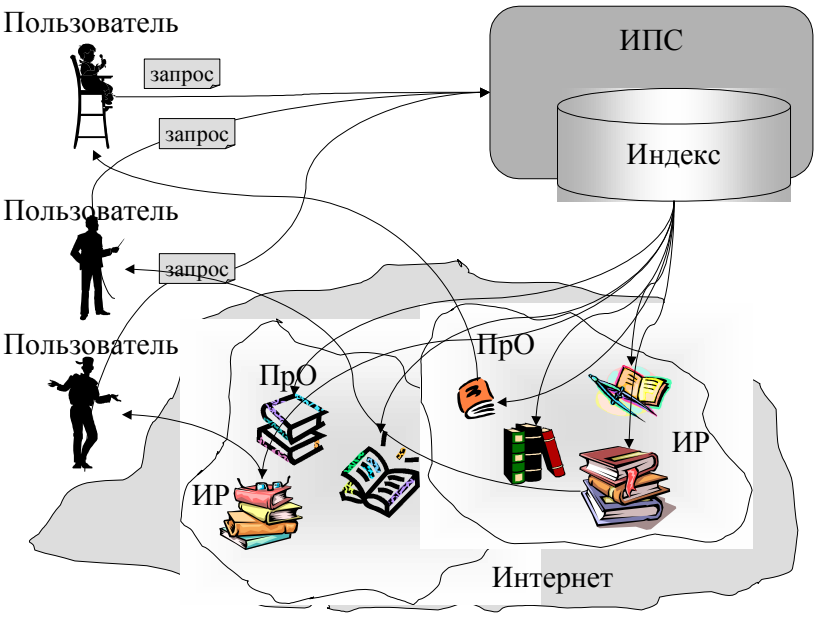

Рис. 2. Взаимодействие пользователей с ИПС

ней категорий через соответствующие темы" и "Онтология – это формальное описание предметной области. Оно определяет общие термины и связи между ними". Очевидно, что второе определение гораздо понятнее, в нем нет специальных терминов, используются более короткие слова и предложения.

Представляется целесообразным использовать для моделирования знаний пользователя о ПрО поиска с помощью частного случая онтологии – тезауруса. Обычно тезаурус Т определяют как словарь, содержащий лексические единицы с явным указанием семантических связей между ними. В данном случае тезаурус определяет словарный запас пользователя и те термины, которыми он может свободно оперировать.

Максимальное количество семантической информации пользователь приобретает при согласовании его тезауруса с содержанием ИР, если информация понятна пользователю и несет сведения, которые отсутствуют в его тезаурусе. Если пользователю предлагают ИР, в котором нет ни одного знакомого ему термина, то он не извлекает из ИР никакой информации. Если же все сведения, содержащиеся в ИР (термины и связи между ними), уже известны пользователю, то никаких изменений в его знаниях тоже не происходит, и, таким образом, семантическое количество информации такого ИР также равно нулю.

Чтобы отфильтровать результаты работы внешней ИПС и получить только те ИР, которые пертинентны информационным потребностям пользователя, необходимо предварительно сформировать тезаурус ПрО, интересующей пользователя, а затем сравнить его с тезаурусами найденных ИР. Будем считать, что тезаурус ПрО – это совокупность терминов, знакомых пользователю ИПС [15]. Это термины, содержащиеся в ИР, найденных ранее по запросам пользователя и признанных им относящимися к этой ПрО [16].

В связи с необходимостью анализа большого количества ИР, мы используем упрощенный алгоритм построения тезауруса: по полному перечню слов, используемых в ИР, строится словарь терминов, из которого отбрасываются стоп-слова, содержащиеся в специально разработанном пользователем списке [17]. Этот алгоритм применяется только для тех ИР, которые не сопровождаются метаописаниями. Иначе из метаописаний (в формате RDF или OWL) извлекаются термины тезауруса и связи между ними, которые дополняют построенный по контенту ИР словарь.

Предлагается использовать оценки удобочитаемости информации для повышения эффективности поиска ИР в Интернет. Для этого можно использовать три пути:

пользователь сам задает желательный для него уровень удобочитаемости текста – формальную удобочитаемость текста $R_{form}(U)$ – для конкретного запроса, а ИПС затем фильтрует полученные результаты выполнения запроса, в первую очередь предлагая пользователю те ИР, оценки удобочитаемости которых ближе всего к желаемой;

уровень удобочитаемости текста, соответствующий образованию пользователя в определенной области, определяется автоматически путем анализа текстовых ИР из соответствующей области, которые пользователь признал понятными и достаточно интересными для него, $R_{form}(U) = \sum_{j=1}^{d} R_{form}(I_j) / d \; j = \overline{1,d}$;

чтобы отделять "длинные" слова ПрО, известные пользователю, от всех остальных "длинных" слов, затрудняющих понимание текста, предлагается использовать тезаурус пользователя, сформированный по понятным пользователю ИР и онтологиям соответствующих ПрО: слова ИР, входящие в тезаурус, обрабатываются в оценках удобочитаемости как короткие независимо от своей реальной длины.

Все три варианта оценки удобочитаемости текста можно применить как ко всему ИР, так и к его аннотации, предлагаемой пользователю в ответ на запрос поисковой системы. В первом случае полученный результат значительно более надежен, но требует закачки на компьютер пользователя всех найденных ИР.

Возникает вопрос, какие из вышерассмотренных оценок удобочитаемости текста следует использовать. Предлагается следующее решение: пользователь отбирает несколько ИР соответствующей ПрО, которые он группирует в одинаковые по удобочитаемости наборы, и для них высчитываются оценки (1)–(11). Тот критерий, который дает наиболее близкие оценки для однотипных ИР и дифференцирует различные по удобочитаемости, и используется в дальнейшем данным пользователем. Такой подход вызван различиями в восприятии текста различными пользователями: для одних сложнее воспринимать текст из длинных запутанных предложений (плохая оперативная память при высокой эрудиции), а для других – со сложными терминами (хорошая оперативная память при низкой эрудиции).

Таким образом, поиск "наилучшего" ИР включает несколько этапов: 1) формирование по набору ключевых слов А – начального множества ИР, релевантных запросу; 2) формирование Т – тезауруса пользователя по отобранным ИР из А и онтологии интересующей его ПрО О; 3) в каждом из ИР в А нужно подсчитать:

- количество терминов из тезауруса пользователя Т и количество связанных с ними терминов в онтологии О;
- среднюю длину предложений и среднее количество слогов в словах в ИР из Х;
- среднюю длину предложений и среднее количество слогов в предложении в ИР из А;

найти ИР из А, наиболее близкие по этим характеристикам к ИР из Х.

Во втором случае следует учитывать ряд особенностей анализируемого текста. Аннотации ИР обычно значительно короче, чем сто слов, и не содержат законченные предложения. Таким образом, для их анализа нельзя применить большинство из рассмотренных критериев. Так как аннотации не содержат законченных предложений, то нельзя оценить среднюю длину предложения. Для простоты подсчета можно считать, что весь анализируемый фрагмент является одним предложением. Кроме того, в аннотации может содержаться информация на различных языках, а коэффициенты для подсчета уровня удобочитаемости в разных языках различаются. Также различны в них и правила разбиения слов на слоги. Поэтому целесообразно оценивать только среднюю длину слов, выраженную в символах. В аннотациях часто присутствуют фамилии авторов

текста с инициалами (одна буква, после которой стоит точка). Если рассматривать инициалы как отдельные слова, то средняя длина слов такого фрагмента будет значительно ниже, чем в аналогичном фрагменте без инициалов. Поэтому предлагается такие слова отбрасывать.

Слова, содержащиеся в тезаурусе пользователя, считаются понятными для него, но несущими большую семантическую нагрузку. Поэтому, независимо от их реальной длины, они заменяются в аннотации строкой "ттт" и рассматриваются в дальнейшем как трехсимвольные. Высчитывается критерий удобочитаемости аннотации (12), являющийся модификацией индекс Колемана–Лиау (5):

$$F_{annot} = 5{,}89 * \frac{x}{k} + 30 * k - 15{,}8, \qquad (12)$$

где x – общее количество символов в строке символов A_{mod}, полученной из аннотации A после отбрасывания инициалов и замены слов тезауруса на "ттт"; k – общее количество слов в A_{mod}.

Программная реализация. Предложенный метод упорядочения найденных ИПС по запросу пользователя ИР на основе подсчета их индивидуальной удобочитаемости используется в интеллектуальной поисковой системе МАИПС [18], ориентированной на пользователей, имеющих постоянные информационные интересы в сети и требующих постоянного поступления новых сведений на интересующую их тему. Функционально МАИПС направлена на выполнение сложных многоразовых запросов в относительно узких областях, связанных с профессиональными или научными интересами пользователей. Запросы таких пользователей могут повторяться от сеанса к сеансу или модифицироваться, но ПрО поиска, в которых пользователи являются экспертами, практически не изменяются.

МАИПС предоставляет пользователям средства для описания интересующих их ПрО через онтологии и тезаурусы. Для того, чтобы создать постоянный запрос, пользователь МАИПС должен: 1) зарегистрироваться в системе, получив свой пароль и логин; 2) из набора предложенных в МАИПС выбрать онтологию, характеризующую интересующую его ПрО; 3) на основе выбранной онтологии сформировать тезаурус интересующей его ПрО (как подмножество терминов онтологии), при необходимости расширив его терминами, отсутствующими в онтологии, но интересными для пользователя; 4) указать желательный уровень удобочитаемости текста искомого ИР (явно или выбрав к качестве образца один или несколько текстовых документов) (рис. 3).

После этого запрос сохраняется в МАИПС как постоянный, и, когда пользо-

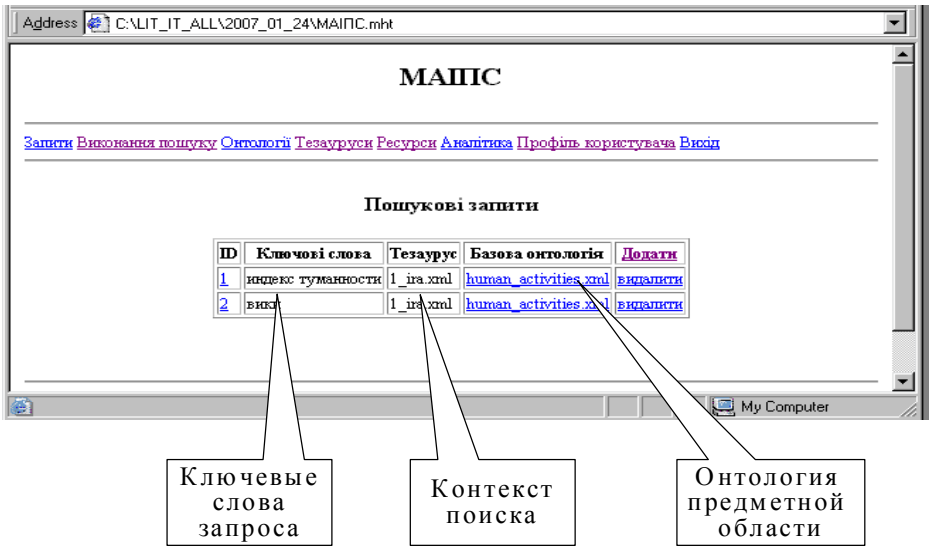

Рис. 3. Создание постоянного запроса в МАИПС

ватель обращается к поисковой системе, введя свой пароль и логин, он может инициировать выполнение этого запроса. Результаты поиска упорядочиваются с учетом наличия в найденных ИР терминов пользовательского тезауруса ("рейтинг ИР") и близости уровня удобочитаемости найденных ИР к заданному пользователем ("индекс читабельности ИР") (рис. 4). Можно, не просматривая ИР, по его индексу читабельности сделать определенные предположения о его пригодности для стоящих перед пользователем целей. Так, проведенные эксперименты показывают, что индекс читабельности научных статей находится обычно в интервале от 18 до 41, технической документации – от 11 до 20, а ИР с индексом читабельности ниже 10, как правило, являются просто набором рекламных слов.

Следует отметить, что для различных запросов, в том числе и соответствующих одному тезаурусу, пользователь может задавать различный уровень удобочитаемости – в соответствии со своей компетентностью в различных подобластях интересующей его ПрО.

В связи с тем, что по запросам пользователей может быть обнаружено много релевантных им ИР, то для более быстрой обработки вначале рейтинг ИР (его соответствие тезаурусу пользователя) и его коэффициент удобочитаемости вычисляются только для аннотации, предлагаемой внешней ИПС. Если ИР заинтересует пользователя, то при нажатии кнопки "Проверить" эти оценки вычисляются с учетом всего текстового контента.

Выводы. Предложенный в работе подход позволяет учитывать формальные характеристики текстовых ИР и соотносить их со знаниями пользователя о предметной области для повышения релевантности поиска и обнаружения на основе критериев удобочитаемости тех информационных ресурсов, которые способны наиболее успешно удовлетворить информационную потребность пользователя, т.е. не только относятся к интересующей пользователя ПрО, но и написаны понятным и интересным для него языком. В современных ИПС такие возможности не реализуются. Предлагается реализовать такие возможности в рамках интеллектуальной ИПС.

Перспективы дальнейших исследований. Сегодня перспективы развития

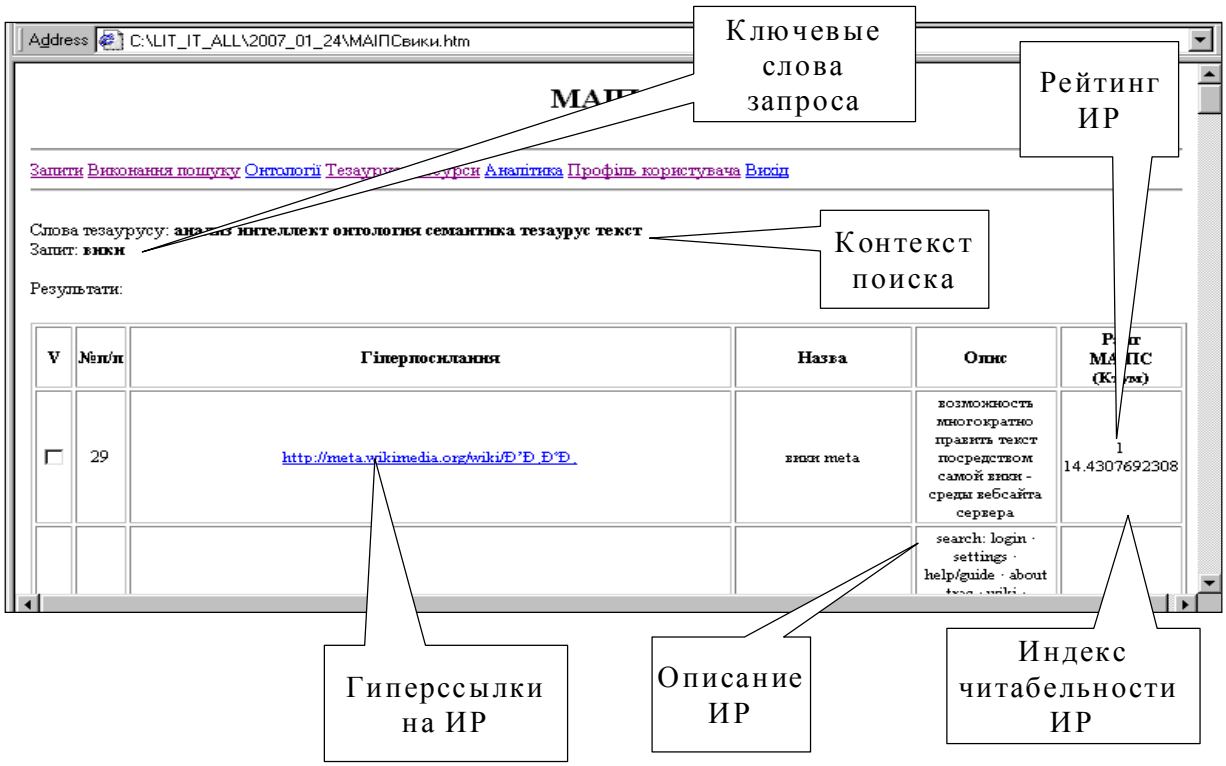

Рис. 4. Результат выполнения запроса к МАИПС

Интернета непосредственно связаны с концепцией Web 2.0., которая ориентирована на создание контента и его классификацию силами не персонала сайтов, а самими пользователями. Web 2.0 позволяет сетевым пользователям одновременно получать информацию из большого количества источников и находить ей новое применение. Например, ИПС Google базируется на методе PageRank, использующем для обеспечения наилучших результатов поиска ссылочную структуру Web, а не характеристики проиндексированных документов. Предлагается для построения индекса удобочитаемости ИР не только анализировать их контент, но и обрабатывать те ресурсы, на которые этот ИР ссылается, так как можно предположить, что в процессе чтения ИР пользователь обратится к этим ссылкам. Какие именно коэффициенты должны использоваться для учета индекса удобочитаемости ссылок, зависит от ПрО поиска и целей, стоящих перед пользователем.

1. *Некрестьянов И., Пантелеева Н.* Системы текстового поиска для Web. – http://meta.math.spbu.ru/~nadejda/papers/web-ir/web-ir.html.
2. *Jansen B.J., Spink A., Saracevic T.* Real life, real users, and real needs: a study and analysis of user queries on the Web T. – http://citeseer.nj.nec.com/jansen00real.html.
3. *Zhu X., Gauch S.* Incorporating quality metrics in centralized/distributed information retrieval on the Worls Wide Web, 2000. – http://citeseer.nj.nec.com/zhu00incorporating.html.
4. *Lawrence S.* Context in the Web Search. – http://citeser.nj.nec.com/lawrence00context.html.
5. *Meacham E.D.* Distance Education: Selecting Textbooks and Writing Study Guides. – http://www.cito.ru/gdenet/technology/print/textbooks/1.
6. *Основные* принципы письменных коммуникаций. – http://dist-cons.ru/modules/Strategy/section2.html.
7. *Miles T.H.* The fog index: a practical readability scale / In Critical Thinking and Writing for Science and Technology. Harcourt Brace Jovanovich. – 1990. – P. 280–284. – http://www.as.wvu.edu/~tmiles/fog.html.
8. *Flesch* Reading Ease Readability Formula. – http://oleandersolutions.com/fleschreadingease.html.
9. *Long A.* Calculating Reading Level. Tameri Guide for Writers. – www.tameri.com/edit/levels.html.
10. *Coleman*–Liau Index http://en.wikipedia.org/wiki/Coleman-Liau_Index.
11. *Raygor* Readability Estimate – http://en.wikipedia.org/wiki/Raygor_Estimate_Graph.
12. *Powers R.D., Sumner W.A., Kearl B.E.* A recalculation of 4 readability formulae // Educational Psychology, University of Birmingham. – 1993. – N 49. – P. 99–105.
13. *McLaughlin H.* SMOG grading – a new readability formula // J. of Reading. – 1969. – N 22. – P. 639–646.
14. *Sticht T.G.* Research towards the design, development and evaluation of a job-functional literacy training program for the US Army // Literacy Discussion, 1973, N 4. – P. 339–369.
15. *Klare G.R.*. The Measurement of Readability. – http://www.timetabler. com/reading.html.
16. *Нариньяни А.С.* Кентавр по имени ТЕОН: Тезаурус + Онтология. – http://www.artint.ru/articles/narin/teon.htm.
17. *Гладун А.Я., Рогушина Ю.В.* Онтологии и мультилингвистические тезаурусы как основа семантического поиска информационных ресурсов Интернет // The Proc. of XII-th Intern. Conf. KDS'2006, Varna, Bulgaria. – 2006. – P.115–121.
18. *Рогушина Ю.В., Гришанова І.Ю.* Засоби інтелектуалізації пошуку мультимедійних даних в Інтернеті // Матеріали Міжнар. наук.-практ. конф. "Розробка систем програмного забезпечення: виклики часу та роль в інформаційному суспільстві". – 2005. – С. 98–101.

Получено 06.12.2006

Об авторе:
Рогушина Юлия Витальевна,
кандидат физико-математичних наук, старший научный сотрудник.

Место работы автора:
Институт программных систем
НАН Украины,
03680, Киев-187,
проспект Академика Глушкова, 40.
тел. (044) 528 4698.