НАВИГАЦИЯ В ПРЕДЕЛАХ ЛЕКСИЧЕСКОЙ ОНТОЛОГИИ С УЧЕТОМ ЧАСТОТНЫХ ФАКТОРОВ[1]

Бодров Д. А., danilb@sphaera.ru
Поляков В. Н., vladimir_polyakov@yahoo.com

Аннотация

В работе предложена когнитивная модель интерфейса к лексической онтологии, основанная на частотных факторах. Предложены стратегии, позволяющие делать частотную маркировку лексической онтологии. В работе рассмотрено влияние трех частотных факторов на модель интерфейса: частотной функции встречаемости слова-узла, веса поддерева, числа подчиненных узлов (лексических термов). Предложены функциональные модели для вычисления этих частотных факторов. Разработан вариант визуализации факторов, оптимально сочетающий их положительные свойства. Проведена апробация предложенной интерфейсной когнитивной модели на специально созданном приложении OntoBrowser. Новизна предлагаемого подхода состоит в использовании частотных факторов для визуализации онтологического дерева. Ожидается, что предложенные решения упростят навигацию для людей с ограниченным знанием языка, подростков, неопытных пользователей. Применение полученных результатов предполагается в проекте интеллектуальной поисковой машины.

Введение

Лексические онтологии, основанные на результатах проекта WordNet [13] являются перспективным для задач поиска информации лингвистическим ресурсом. Этот тип представления лексикографических данных уже нашел свое применение в задачах навигации в среде WWW [3,4,12], активно применяется в задачах переформулирования и вербализации запросов в поисковых системах [1,5,11,14,15,21,24,26,29]. Онтологии успешно используется в задаче разрешения лексической многозначности [2,7,8,9,10,16,19,20,25,27], важной с точки зрения релевантности поиска.

Как справедливо отмечалось в [17,18], лексические онтологии имеют много общего с тезаурусами, что вызывает некоторые сомнения в возможности рассматривать онтологии как лексикографический ресурс нового типа. В связи с этим отметим, что они имеют один семантический признак, отчетливо определяющий их самостоятельный статус. В онтологиях лексическое значение постулировано в явном виде, в то время как в тезаурусах оно не определяется вообще или может быть выделено косвенно на основе совокупности межлексемных связей. Для постулирования значений в лексических онтологиях применяется несколько способов, которые обычно используются в совокупности: указание толкования значения в явном виде, указание синсета[2], привязка к онтологическому классу, ссылка на набор примеров из корпуса текста.

Организация удобной навигации по БД лексической онтологии представляет собой самостоятельную задачу, от решения которой во многом зависит успешное решение проблемы информационного поиска. Имеется несколько завершенных проектов по организации удобного интерфейса к лексическим онтологиям.

См. например страницы:

«WordNet. Related Projects»

(http://www.cogsci.princeton.edu/~wn/links.shtml)

Greg Peterson. WordNet Vocabulary Helper

[1] Данная работа выполнена при частичной финансовой поддержке фонда РФФИ (грант 02-0790413).
[2] Синсет – множество синономов. Термин, введенный в рамках проекта WordNet.

(http://poets.notredame.ac.jp/cgi-bin/wn)

Nikolai Golovchenko. Web interface for WordNet 1.6

(http://www.sxlist.com/cgi-bin/wnsearch.exe)

Однако нельзя сказать, что предложенные интерфейсы для навигации по онтологиям являются до конца интуитивными и легкими в использовании для всех категорий пользователей. Особенно большие затруднения могут возникать при использовании онтологической сети категорией пользователей, для которой язык лексической онтологии не является родным. Похожие затруднения могут также испытывать люди, имеющие ограниченный словарный запас. Это в первую очередь люди с недостаточным знанием компьютерных технологий и навыками работы с информационными системами. Связано это со следующими факторами:

- присутствие в сети WordNet технических уровней;
- непрозрачность уровней онтологий;
- отсутствие дополнительных средств фокусировки и фильтрации лексических узлов[3];
- неполное знание иностранной и специальной лексики.

В то же время, учитывая все более широкое распространение поисковых ресурсов в сети WWW, можно предположить, что совершенствование интерфейса к лексическим онтологиям с ориентацией на эти очень большие категории пользователей имеет практический, экономический и научный интерес.

Одним из способов применения онтологических сетей может являться их использование в качестве средств фокусировки и расширения запросов для поисковых машин [22].

Другое применение онтологий — это вербализация запросов в ситуации, когда пользователю хорошо известен онтологический класс информационного объекта, однако нет информации о том, как лексикализованы его название и характеристические свойства. По нашему мнению, режим вербализации запроса с применением онтологий может быть плодотворно применен в сфере патентного и научного поиска.

В настоящей работе предложен подход к организации интерфейса к лексической онтологии типа WordNet, основанный на частотных зависимостях встречаемости слов в тексте. Кроме того, в статье предложено два дополнительных интерфейсных решения, также основанных на частотных факторах: сортировка термов по частоте запросов к поисковой машине и сортировка термов по частоте их активации в процессе поиска по онтологическому дереву.

- В качестве базы онтологической сети были использованы данные проекта WordNet (версия 1.6.), инициированного в Принстонском университете (Princeton University, New Jersey, USA) для английского языка в 1985 г. и успешно развивающегося до сих пор [13].

В работе для целей навигации предложено использовать три частотных фактора: частотная функция встречаемости слова-узла, вес поддерева, число подчиненных узлов (лексических термов). Было предложено несколько моделей учета частотных факторов при организации интерфейса. Для апробации моделей было разработано приложение-прототип OntoBrowser. Целью этой разработки является только демонстрация интерфейса к онтологической сети с учетом частотного фактора, но не решение фундаментальных задач обработки текстов на естественном языке. Также в статье уделено внимание принципам разработки когнитивно-ориентированного интерфейса на базе частотных факторов. Результаты работы планируется использовать в рамках проекта Интеллектуальная поисковая машина [23].

Использование онтологий для переформулирования и вербализации запросов в поисковых системах

Общая идея фокусировки запроса и расширения поиска заключается в том, что пользователь не всегда способен с первого раза сформулировать запрос с удовлетворяющей его степенью качества. В этом случае онтологии выступают средством переформулирования запроса, обеспечивающим достижение результата, т.е. необходимого качества запроса.

Определим пертинентность P как процент документов, удовлетворяющих информационным потребностям пользователя поисковой системы:

[3] Имеется в виду дополнительные средства фокусировки, а не штатные, связанные с родовидовыми и подобными им связями (меронимия и т. д.).

$$P = \frac{N_p}{N} \cdot 100 \quad , \qquad (1)$$

где P — пертинентность;

N_p — число документов, удовлетворяющих информационную потребность,

N — общее число документов.

Если запрос привел к низкому уровню пертинентности, и высокому уровню информационного шума применяется операция фокусировки. Фокусировка запроса — это операция переформулирования запроса, которая позволяет повысить пертинентность. Наиболее типичным способом фокусировки с помощью лексических онтологий является выбор гипонима в качестве ключевого слова.

Расширение поиска — это операция переформулирования запроса, которая применяется в случае отрицательного результата поиска или когда найдено слишком мало документов.

Типичным способом расширение поиска с помощью лексических онтологий является выбор когипонимов и/или гиперонима. Эти механизмы перефразирования запросов являются частью интерфейсного решения в проекте ИПМ [23]. Разумеется, что для применения указанных механизмов поисковая машина должна включать корпус текстов, семантически размеченный по узлам лексической онтологии.

Рассмотрим пример, когда пользователю необходимо найти фамилии знаменитых футболистов. Если он не американец, то он может не знать, что европейскому названию *football* на североамериканском континенте соответствует названию *soccer*, из-за чего может возникнуть ошибка перевода.

Person -> contastenant -> athlete -> football player -> ???

Онтологическая связь от слова *athlete* поможет восстановить правильную цепочку поиска.

Person -> contastenant -> athlete -> soccer player -> goalkeeper -> Lev Yashin

Существует область применения поиска и навигации по онтологиям, которая потенциально может дать гораздо более весомые аргументы в их пользу. Речь идет о применении онтологий для патентного поиска. При поиске аналогов нового изобретения чрезвычайно трудно сформулировать название объекта. Тем более, невозможно предсказать всевозможные имена собственные, которые были использованы в качестве торговых марок образцов-аналогов или взяты по именам авторов изобретений. В этом случае онтология позволяет локализовать поиск определенным классом объектов, внутри которого уже можно осуществлять более детальный просмотр документов. Например, изобретатель нового типа двигателя может начать свой просмотр подходящих патентов, начиная с концепта *engine* (*sense* 1):

Engine *-- (motor that converts thermal energy to mechanical work)*

=> aircraft engine -- (the engine that powers and aircraft)

=> automobile engine -- (the engine that propels an automobile)

=> auxiliary engine, donkey engine -- (a small engine (as one used on board ships to operate a windlass))

=> generator -- (engine that converts mechanical energy into electrical energy by electromagnetic induction)

=> heat engine -- (any engine that makes use of heat to do work)

=> reaction-propulsion engine, reaction engine -- (a jet or rocket engine based on a form of aerodynamic propulsion in which the vehicle emits a high-speed stream)

Далее, предположим, что речь идет о тепловых двигателях.

heat engine *-- (any engine that makes use of heat to do work)*

=> external-combustion engine -- (a heat engine in which ignition occurs outside the chamber (cylinder or turbine) in which heat is converted to mechanical energy)

=> internal-combustion engine, ICE -- (a heat engine in which combustion occurs inside the engine rather than in a separate furnace; heat expands a gas that either moves a piston or turns a gas turbine)

И если это — двигатель внутреннего сгорания, онтология приведет пользователя к более узкому классу конструкций двигателей, среди которых уже можно осуществлять просмотр патентов.

internal-combustion engine, ICE *-- (a heat engine in which combustion occurs inside the engine rather than in a separate furnace; heat expands a gas that either moves a piston or turns a gas turbine)*

=> diesel, diesel engine, diesel motor -- (an internal-combustion engine that burns heavy oil)

=> four-stroke engine, four-stroke internal-combustion engine -- (an internal-combustion engine in which an explosive mixture is drawn into the cylinder on the first stroke and is compressed and ignited on the second stroke; work is done on the third stroke and the products of combustion are exhausted on the fourth stroke)

=> gas engine -- (an internal-combustion engine similar to a gasoline engine but using natural gas instead of gasoline vapor)

=> gasoline engine -- (an internal-combustion engine that burns gasoline; most automobiles are driven by gasoline engines)

=> outboard motor, outboard -- (internal-combustion engine that mounts at stern of small boat)

=> radial engine, rotary engine -- (an internal-combustion engine having cylinders arranged radially around a central crankcase)

=> reciprocating engine -- (an internal-combustion engine in which the crankshaft is turned by pistons moving up and down in cylinders)

=> rotary engine -- (an internal-combustion engine in which power is transmitted directly to rotating components)

=> valve-in-head engine -- (internal-combustion engine having both inlet and exhaust valves located in the cylinder head)

Частотный анализ корпуса запросов к поисковой системе Yandex, выполненный в [23] показал, что свыше 90 процентов запросов в тексте содержат имена существительные или именные группы. Это позволяет при навигации по онтологии ограничиться только именами существительными. В разработанной авторами исследовательской прикладной программе OntoBrowser имеется фильтр по частям речи, позволяющий осуществлять такую возможность.

Частотные факторы

Организация лингвистические исследований, основанных на частотных зависимостях, являются весьма распространенным с прагматической точки зрения подходом, позволяющим сэкономить исследовательские ресурсы и одновременно охватить подавляющую часть явлений языка.

Базовая посылка настоящего исследования заключается в том, что в силу частотных закономерностей большинство пользователей поисковых систем интересует именно частотная лексика. При этом остается открытым вопрос, какие частотные зависимости являются полезными при организации интерфейса по лексическим онтологиям? Авторам представляется, что на поставленный вопрос нет однозначного ответа.

Можно выделить следующие четыре гипотезы, позволяющие организовать частотно-зависимый онтологический интерфейс для различных категорий и информационных потребностей пользователей:

1. Маркирование лексики с максимальной частотой использования.

Можно предположить, что существует категория пользователей, которую не интересует специфическая и редкая, а вполне удовлетворяет общеупотребительная в рамках данного онтологического класса лексика. К такой группе относятся категории пользователей с ограниченным словарным запасом, о которых мы говорили выше.

2. Маркирование лексики с минимальной частотой использования.

Некоторых пользователей не интересует информация, связанная с широкоупотребительной лексикой, а интересуют достаточно специфические слова. К этой категории относятся научный и инженерный персонал фирм и организаций.

3. Маркирование лексики, наиболее часто встречающейся в запросах других пользователей к поисковой системе.

Этот вид частотной зависимости характерен для большинства постоянных пользователей при выборе поисковых тем в Интернете. В данном случае оправданием такой организации интерфейса служат объективные закономерности в «потреблении» информации, основанные на частотном анализе запросов к поисковым системам. Пример частотного анализа запросов можно найти в работе [6].

4. Маркирование наиболее часто используемых путей в лексической онтологии.

Блуждание в лексической онтологии можно уподобить навигации в WWW. Таким образом, каждый пользователь оставляет в онтологии свой след в виде пути от стартового узла до того места в сети, в котором он формулирует запрос к поисковой системе. Наделив узлы на пути навигации пользователей счетчиками, фиксирующими количество проходов, можно получить очень интересные частотные зависимости.

Все эти гипотезы можно рассмотреть с точки зрения уровня языковой компетенции. При такой постановке вопроса, очевидно, что между частотой лексики и словарным запасом пользователя должна быть обратная зависимость, т. е. чем меньше словарный запас, тем более общеупотребительная лексика и действия других пользователей его интересуют.

Используя частотную информацию, опирающуюся на одну из этих гипотез, можно организовать навигацию по онтологиям более эффективным способом, при этом главным показателем эффективности является экономия времени на поиск нужного слова при переформулировании неудачного поискового запроса.

Для отображения частотного фактора используется три показателя:

- частотная функция узла;
- вес поддерева;
- число подчиненных частотных узлов.

Частотная функция узла

В нашей модели под частотной функцией узла понимается величина, зависящая от частоты встречаемости связанного с узлом словарного терма в тексте на тысячу слов[4]. Эта величина в общем случае является функцией от закономерностей распределения частот всех слов словаря и может быть вычислена разными способами.

Необходимо заметить, что сама зависимость f_i на диапазоне лексических термов носит гиперболический характер (закон Ципфа [28]). Однако авторам неизвестны когнитивные исследования, в которых бы имелся ответ на вопрос: Как зависит восприятие термина человеком от частотности этого термина в текстах? Кроме того, немаловажную роль может сыграть и специализация пользователя на текстах определенной тематики. В качестве рабочей гипотезы будем считать, что такая зависимость имеется, однако нет однозначного подтверждения, что она линейная. Поэтому в предлагаемой интерфейсной модели предусмотрена возможность функционального преобразования для усиления или сглаживания изначального частотного распределения.

В зависимости от конечной цели построения частотного интерфейса можно сконструировать различные частотные функции, которые должны обладать следующим основным свойством:

$$x >= y \Rightarrow F(x) >= F(y) \qquad (2)$$

и дополнительными свойствами (при их отсутствии требуется дополнительная нормировка значений частотных

[4] Под словарным термом понимается слово или устойчивое словосочетание (коллокация). Нормировка частоты на 1000 слов является общепринятым приемом в частотных лингвистических исследованиях.

функций):

$$x \in [0, 1] \Rightarrow F(x) \in [0, 1] \quad (3)$$

$$F(0) = 0,\ F(1) = 1 \quad (4)$$

В эксперименте нами были использованы описанные ниже частотные функции.

Линейная шкала:

$$F_i = f_i, \quad (5)$$

где F_i — частотная функция узла, используемая для индикации; f_i — нормированная частота узла.

Под нормированной частотой узла понимается относительная частота узла, нормированная тем или иным способом так, чтобы значения лежали в интервале [0;1]. При этом для достижения наилучших результатов предпочтительно, чтобы значения 0 и 1 нормированных частот соответствовали минимальной и максимальной частоте выборки. Это позволит максимально полно использовать индикаторы и более эффективно визуализировать диапазон частот.

Здесь и далее мы использовали следующее правило нормировки:

$$V_N = \begin{cases} \dfrac{V - V_{\min}}{V_{\max} - V_{\min}}, V_{\max} <> V_{\min} \\ 1, V_{\max} = V_{\min} \end{cases}, \quad (6)$$

где V_N — нормированное значение;

V — исходное значение;

$V_{\min}$ — минимальная частота выборки;

$V_{\max}$ — максимальна частота выборки.

Вырожденный случай, когда максимальное и минимальное значение выборки совпадают, в контексте рассматриваемой проблемы подразумевает, что все частоты равны, а следовательно не представляет для нас интереса. В таком случае нормированные значения подразумеваются равными 1.

Квадратичная «усиливающая» шкала:

$$F_i = f_i^{\,2} \quad (7)$$

Корневая «сглаживающая» шкала:

$$F_i = \sqrt{f_i} \quad (8)$$

На рис. 1 проиллюстрировано усиливающее и сглаживающее действие степенных шкал по сравнению с линейной. Интерфейс отображает значения линейной шкалы. Средняя кривая — аппроксимацию рассчитанных значений линейной частотной функции гиперболической функцией, нижняя — квадратичной функции, верхняя — корневой функции.

В то время, как квадратичная шкала увеличивает разрыв между значениями близкими к максимальному и минимальному, корневая его уменьшает. Руководствуясь желанием сокращения или увеличения этого разрыва и степенью этих изменений можно сконструировать и другие частотные функции[5].

[5] Например, можно рассмотреть более общие случаи этих шкал: $F_i = f_i^{\,m}$ и $F_i = \sqrt[m]{f_i}$ или различные логарифмические преобразования.

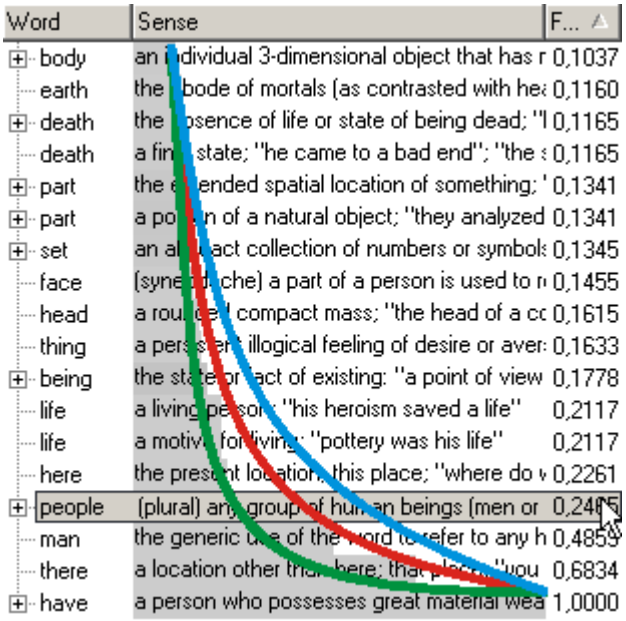

Рис. 1. Влияния вида частотной функции на интерфейс

При вычислении частотных функций узлов в зависимости от целей построения интерфейса возможны два различных подхода. В первом случае нормализация частот происходит для всего словаря. Такой вариант более предпочтителен для различных исследовательских целей на всей онтологической сети. В другом варианте расчет производится в пределах одного онтологического класса (от вершины дерева вниз). Такой вариант применим практически в различных поисковых и других системах, где одновременно отображается не более одной ветви. Именно так выглядит интерфейс при использовании лексической онтологии для целей переформулирования запроса. В таком случае пользователю предоставляется оценка значимости слов в пределах интересующего его класса, а высокочастотные (в корпусе текстов) слова, не попадающие в рассматриваемый класс (а значит представляющие собой в данном контексте шум), исключаются из расчетов. Применение расчета частотного распределения в пределах онтологического класса оправдано также тем, что для профессионала распределение специальной лексики по важности оказывается скорее всего иным, чем для остальных и, как представляется, связано с частотным распределением внутри онтокласса.

Вес поддерева

Под весом поддерева подразумевается величина, зависящая от частотных факторов входящих в него дочерних узлов (гипонимов). Подобно частотным функциям, можно сконструировать множество методов вычисления весов. Единственным жестким требование к этим методам, является прямая (а не обратная) зависимость от частот входящих узлов.

Условно методы можно разделить на два класса:

- не учитывающие частоту корневого узла;
- учитывающие частоту корневого узла.

На практике целесообразнее применять первую группу, т. к. во второй группе можно получить достаточно высокий вес поддерева при весьма низких частотах входящих узлов. Таким образом, существует риск указать пользователю нецелесообразный путь с неиспользуемыми словами.

Рассчитанные веса узлов могут быть (в зависимости от используемого метода) слишком маленькими (с диапазоном меньше «цены деления» индикатора) или выходить за пределы интервала [0; 1] (например, для аддитивных методов). Однако для отображения весов в когнитивно-ориентированном интерфейсе целесообразно добиться распределения показателей в интервале [0; 1], для чего следует нормализовать полученные значения весов, используя формулу (6). Подобно частотным факторам узлов, нормализацию весов также можно произвести по всему множеству узлов или в пределах онтологического класса. По причинам, отмеченным выше, второй вариант целесообразнее.

Далее приводятся несколько вариантов методов.

Усредненный:

$$W_j = \frac{\sum_{i=1}^{n} F_{ji}}{n}, \qquad (9)$$

где W_j — вес поддерева;
n — число дочерних узлов;

F_{ji} — частотная функция дочернего узла *i*.

Основной недостаток этого метода заключается в «размывании» результата за счет служебных (технических) уровней и слабочастотных слов.

Усредненный с отсечением:

$$W_j = \frac{\sum_{i=1,n,F_{ji}>F_0} F_{ji}}{n}, \qquad (10)$$

где F_0 — заданный уровень отсечения.

Данный метод избавлен от недостатка предыдущего за счет исключения низкочастотных узлов, однако может искажать результат.

Эти методы (усредненный и усредненный с отсечением) можно легко модифицировать для получения методов второго типа, добавив в расчет корневой узел. Учитывая тривиальность изменений, формулу здесь не приводим.

Максимальный из подчиненных:

$$W_j = \max_{i=1,n}(F_{ji}) \qquad (11)$$

Данный метод не вполне точно характеризует вес поддерева, т. к. не учитывает число узлов, однако в этом случае частотная функция высокочастотного узла служит своего рода маяком при навигации.

Максимальный из подчиненных уровнем ниже:

$$W_j = \max_{i=1,n,l_{ji}=1} (F_{ji}), \qquad (12)$$

где l_{ji} — уровень узла (расстояние от узла до вершины поддерева).

Метод позволяет экономить ресурсы, ограничивая количество просматриваемых узлов, однако к недостатку предыдущего метода добавляет опасность получить неоправданно низкий вес поддерева в случае большого числа низкочастотных или технических узлов на следующем уровне.

Комбинированный:

$$W^*_i = F_i + W_i, \qquad (13)$$

где F_i — вес узла, порождающего поддерево;
W_i — вес поддерева, рассчитанный одним из предыдущих методов.

В отличие от предыдущих методов, где нормирование желательно для получения более наглядных индикаторов, для последнего метода нормирование обязательно, т. к. существует вероятность получения значения, не лежащего в диапазоне [0; 1].

Вычисление показателей веса поддерева и предоставление этой информации пользователю позволяет оценивать

перспективность направлений навигации по онтологической сети без углубления на каждом из узлов, что также сокращает время переформулирования запроса.

Число подчиненных частотных узлов

Число подчиненных частотных узлов N_i показывает число «перспективных» узлов в поддереве.

$$N_i = \left| \{ L_i, W_i > W_0 \} \right|, \qquad (14)$$

N_i определяется как мощность множества L лексических термов онтологии, для которых вес поддерева W_i превышает пороговое значение W_0.

Аналогичный частотный фактор можно сконструировать на базе частотной функции узла:

$$N_i = \left| \{ L_i, F_i > F_0 \} \right|. \qquad (15)$$

Этот метод не учитывает общего количества узлов или величину отклонения от порогового значения.

Пороговую частоту веса узла можно использовать как частоту отсечения малочастотных узлов в случае принятия гипотезы о том, что пользователей интересует более употребительная лексика. Возможны два варианта отсечения:

- удаление из сети узлов с низким весом, но сохранением их потомков;
- удаление поддерева целиком в случае, если вес поддерева ниже некоторой пороговой величины;
- выделение в отдельном списке заголовков частотных узлов.

Первый и второй вариант подразумевают исключение части информации из интерфейса, кроме того второй вариант несет в себе опасность удаления весьма важных узлов, т. к. в некоторых больших поддеревьях высокую частоту могут имеют лишь несколько узлов, что при использовании большинства методов расчета веса даст низкий общий вес поддерева, а следовательно оно будет отсечено.

Так как при проведении исследования потеря части информации представлялась нежелательной, для реализации в тестовом приложении был выбран третий вариант.

Также отсечение можно организовать путем задания не порога частоты, а доли от общего числа отсекаемых узлов с малой частотой.

Способы визуализации частотных факторов

При выборе способов визуализации частотных факторов мы исходили из следующих соображений:

- метод должен быть интуитивно-понятным и не требующим дополнительных разъяснений;
- необходимо компактное и в тоже время полное отображение информации;
- необходимо органичное сочетание способов визуализации с характером информации, представленной в лексических онтологиях и между собой.

Основная цель средств облегчения навигации — указать пользователю наиболее эффективный путь. Однако ввиду того, что онтологическая сеть и без того сильно насыщена информацией, вынесение в интерфейс дополнительных данных может сыграть как положительную, так и отрицательную роль из-за рассеивания внимания пользователя. Следовательно, при разработке когнитивной модели интерфейса для визуализации частотных факторов необходимо было найти способ представления информации, не требующий отдельной концентрации на ней внимания пользователя.

В итоге были рассмотрены и опробованы на практике следующие средства визуализации частотных факторов:

- цвет шрифта;
- насыщенность цвета шрифта;
- размер шрифта;
- линейный индикатор;

- цифровые индикаторы.

В результате экспериментирования с разными комбинациями факторов мы пришли к выводу, что оптимальным сочетанием визуальных факторов будет следующее:

- для отображения частотного фактора лучше всего подходит линейный индикатор (длина индикатора отображает относительную частота узла, насыщенность тона — вес поддерева);
- для отображения порога отсечения лучше подходит сочетание насыщенности и размера шрифта;
- для отображения числа узлов выше уровня отсечения лучше всего подходит цифровой индикатор.

Эти выводы основаны на следующих соображениях когнитивного характера.

Частота и длина линейного индикатора имеют хорошую ассоциативную связь, кроме того, линейный индикатор используется традиционно в различных пользовательских интерфейсах (например, в интерфейсах поисковых машин). Использование линейных индикаторов дает наиболее высокую точность (после цифровых данных) представления информации, т. к. при использовании остальных вариантов либо трудно увидеть различия между близкими уровнями (для насыщенности) либо шкала имеет слишком маленький диапазон вариации (возможность изменения размера шрифта ограничена несколькими пунктами). Также в силу расположения линейных индикаторов вертикально друг под другом пользователь имеет возможность легкого и безошибочного сопоставления показателей нескольких узлов.

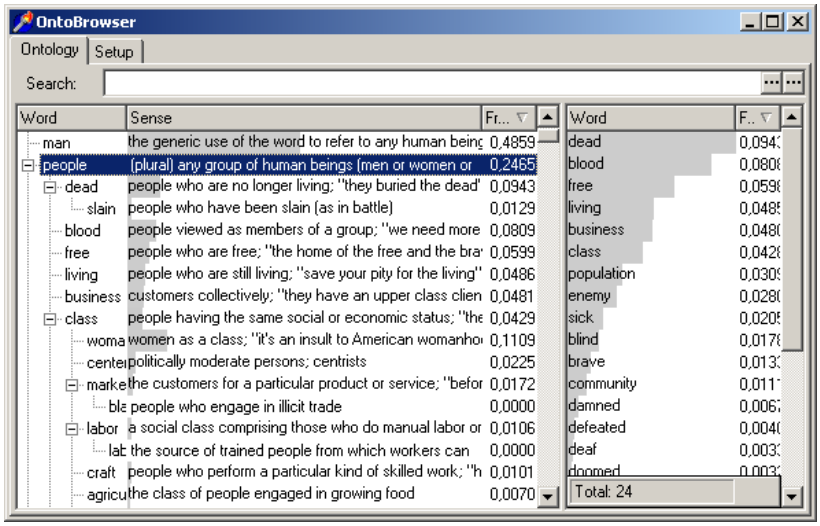

Рис. 3. Общий вид приложения OntoBrowser

Вариация размера и цвета шрифта создает эффект фокуса внимания и периферии, а увеличенный размер дает дополнительный эффект приближения слова к пользователю, что также вводит слово в фокус. Однако диапазон изменения размеров шрифта не должен быть больше 1–3 пунктов, т. к. иначе возникает впечатление неаккуратности интерфейса, что ограничивает применимость данного подхода.

Эксперименты по использованию цвета в качестве индикатора частоты или веса показали, что цветовая раскраска с привязкой к частоте вводит пользователя в заблуждение в силу отсутствия оппозиционного когнитивного представления о цветах на цветовой шкале и требует дополнительных разъяснений о привязке цвета к величине частоты. Использование монохроматической шкалы дало лучшие результаты. Насыщенность цвета полоски наталкивает пользователя интуитивно на мысль о том, что плотность информации в этом поддереве выше, чем в других.

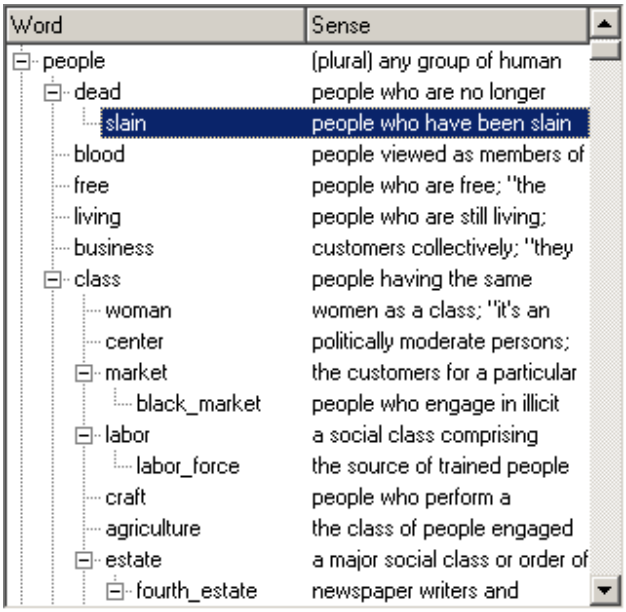

Рис. 3. Онтология без визуализации частотного фактора

Дополнительную информацию, например число узлов выше уровня отсечения, лучше всего отображать с помощью цифровых индикаторов. Из-за обилия когнитивных элементов интерфейса такая информация, являющаяся второстепенной при навигации, не будет бросаться в глаза и дополнительно нагружать и так уже несколько перенасыщенный информацией интерфейс.

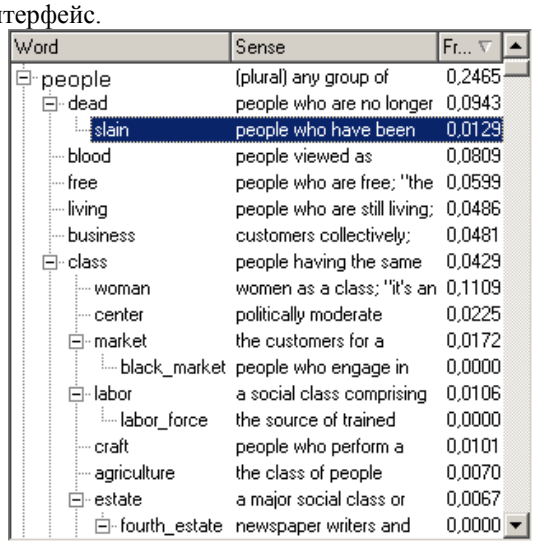

Рис. 4. Визуализации частотного фактора размером шрифта

Апробация предложенной интерфейсной когнитивной модели проводилось в приложении OntoBrowser. Общий вид приложения представлен на рис. 3. В левом окне отображается онтологическое дерево с визуализаций частотного фактора, в правом — частотные узлы (с частотой выше частоты отсечения) выбранного поддерева и их количество.

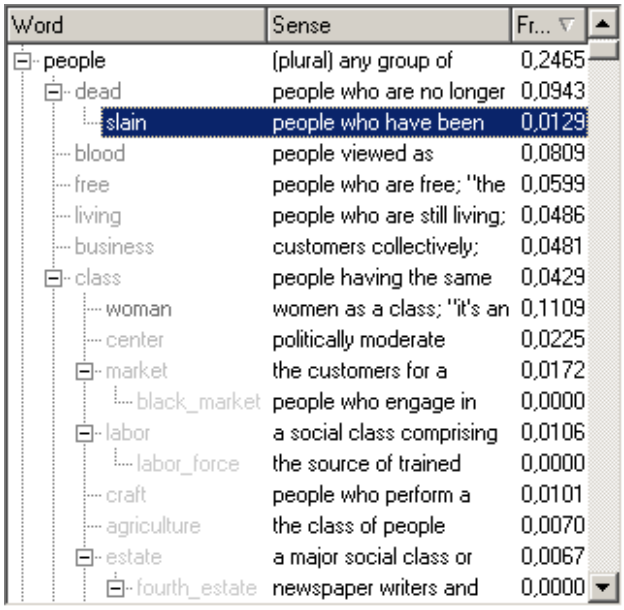

Рис. 5. Визуализации частотного фактора насыщенностью тона

На рис. 3–7 изображены примеры визуализации частотных факторов в программе. В качестве отправной точки для тестирования нами была выбрана гипотеза о полезности частотной лексики при навигации по онтологии. Для проведения частотных исследований анализировался набор текстов различных стилей и тематики на английском языке объемом 17 189 574 слов. Анализ проводился в два этапа. На первом этапе было проведено составление полного словаря корпуса текстов с указанием абсолютной частоты слов. На втором этапе этот словарь сопоставлялся с базой данных WordNet 1.6 и проводилось вычисление относительных частот сопоставленных слов.

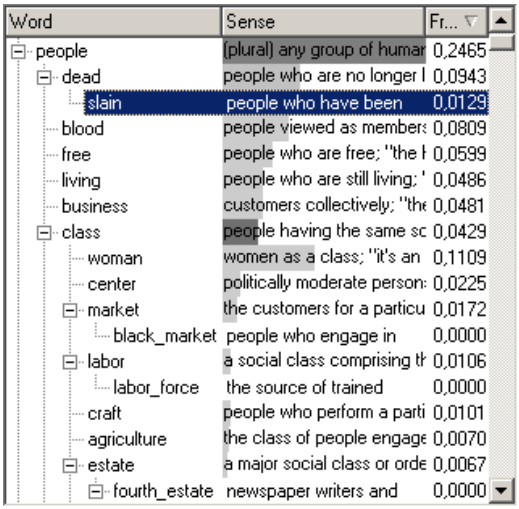

Рис. 6. Визуализации частотного фактора (линейная шкала) и веса поддерева (максимальный из подчиненных) линейным индикатором

Также на рис. 3–7 показаны примеры отображения онтологии с различными способами визуализации значения частотной функции словарных термов (F_i).

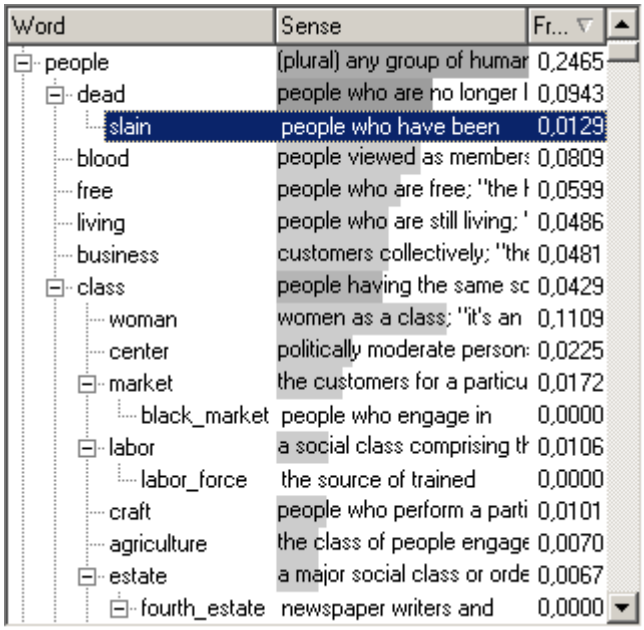

Рис. 7. Визуализации частотного фактора (корневая шкала) и веса поддерева (усредненный) линейным индикатором

Из примеров видно, что использование частотных факторов способно значительно облегчить процесс навигации по онтологии.

Обсуждение результатов

При верификации полученных результатов мы рассматривали возможные неучтенные факторы. Ограничения использованной нами методики частотного анализа сводятся к следующим пунктам:

- при расчете частот не учитывался фактор многозначности, т.е. частота считалась по лексеме, а не по ее значению;
- в процессе частотного анализа не учитывалась морфология, т. е. считались только слова в нормальной форме;
- была использована сравнительно небольшая выборка текстов[6].

Однако эти ограничения можно считать не слишком существенными для целей нашего исследования, т. к. они не снимают частотных зависимостей, а только сглаживают их[7].

Как показали наши экспериментальные исследования, эффективность применения предложенных частотных моделей в значительной степени определяется учетом когнитивных факторов при организации интерфейса. Чтобы подчеркнуть этот аспект нашего исследования, перечислим еще раз эти факторы. К ним относятся:

- расчет частотных функций узлов в пределах выбранного класса;
- нормировка весовых и частотных функций термов в интервале [0,1];
- органичное сочетание средств отображения частотных факторов (линейный индикатор, насыщенность цвета и размер шрифта, цифровой индикатор).

Заключение

В работе была предложена когнитивная модель интерфейса к лексической онтологии, основанная на предположении о влиянии частотности лексики на результаты поиска. Рассмотрение частотных факторов осуществлялось для двух поисковых задач, в которых целесообразно применение лексических онтологий: задача переформулирования запроса и задача вербализации запроса для патентного поиска. Были сформулированы четыре

[6] Пересечение словников выбранного корпуса текстов и системы WordNet составило 44 процента, что, на наш взгляд, вполне достаточно для валидности результатов исследования.

[7] За исключением фактора многозначности, который оказывает более сложное влияние, однако на меняет общих принципов организации когнитивно-ориентированного интерфейса, основанного на частотных зависимостях.

частотные стратегии, позволяющие делать частотную маркировку лексической онтологии и высказаны предположения о группах пользователей, для которых каждая из стратегий является подходящей. В работе рассмотрено влияние трех частотных факторов на модель интерфейса: частотной функция встречаемости слова-узла, веса поддерева, числа подчиненных узлов (лексических термов). Рассмотрены различные модели для вычисления этих частотных факторов. Исследованы различные варианты их визуализации и предложен вариант, оптимально сочетающий эти факторы. Проведена апробация предложенной интерфейсной когнитивной модели на специально созданном приложении OntoBrowser. Применение полученных результатов предполагается в проекте интеллектуальной поисковой машины. Показано, что применение частотных факторов для фильтрации и структурирования онтологического пространства может дать положительный эффект.

Благодарности

Авторы выражают свою признательность сотруднице Школы бизнеса г.Копенгаген (Дания) Екатерине Мхаанна за ряд ценных замечаний, сделанных по работе.

Список использованных источников

1. T. Andreasen,, J. Fischer Nilsson, & H. Erdman Thomsen: Ontology-based Querying, in H.L. Larsem et al. (eds.) Flexible Query Answering Systems, Flexible Query Answering Systems, Recent Advances, Physica-Verlag, Springer, 2000. pp. 15-26.
2. Banerjee, Satanjeev and Ted Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02)*, Mexico City, February, 2002.
3. Brezeale, Darin. *The Organization of Internet Web Pages Using WordNet and Self-Organizing Maps.* Masters thesis, University of Texas at Arlington, August 1999.
4. Chakravarthy, A. S. and K. B. Haase. NetSerf: using semantic knowledge to find Internet information. In: *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, 1995.
5. Gonzalo, Julio, Felisa Verdejo, Irina Chugur and Juan Cigarran. Indexing with WordNet synsets can improve text retrieval. In: *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.
6. Jansen, B. J., Spink, A., and Saracevic, Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management. 36(2), T. 2000. 207-227.
7. Karov, Yael and Shimon Edelman. Learning similarity-based word sense disambiguation from sparse data. In: *Proceedings of the 4th Workshop on Very Large Corpora*, Copenhagen, 1996.
8. Kwong, Oi Yee. Word sense disambiguation with an integrated lexical resource. In: *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.
9. Li, Xiaobin, Stan Szpakowicz and Stan Matwin. A WordNet-based algorithm for word sense disambiguation. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995, pp. 1368 - 1374.
10. Lin, Dekang. Using syntactic dependency as local context to resolve word sense ambiguity. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, 1997.
11. Mandala, Rila, Tokunaga Takenobu and Tanaka Hozumi. The use of WordNet in information retrieval. In: *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.
12. Martin, Philippe. Using the WordNet Concept Catalog and a Relation Hierarchy for Knowledge Acquisition In: Proceedings of Peirce'95, 4th International Workshop on Peirce, University of California, Santa Cruz, August 1995.
13. Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. Introduction to WordNet: an on-line lexical database. In: *International Journal of Lexicography* 3 (4), 1990, pp. 235 - 244.
14. Mihalcea, Rada and Dan I. Moldovan. eXtended WordNet: progress report. In: *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.
15. Mihalcea, Rada and Dan I. Moldovan. A WordNet-Based Interface to Internet Search Engines In: *Proceedings of FLAIRS-98*, May 1998, Sanibel Island, FL.
16. Mihalcea, Rada and Dan I. Moldovan. Word sense disambiguation based on semantic density. In: *Proceedings of the*

COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.

17. Нариньяни А.С. Кентавр по имени ТЕОН. Тезаурус+онтология. Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т.1. с.184-188.
18. Нариньяни А.С. ТЕОН-2. От тезауруса к онтологии и обратно. Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. (Протвино, 6-11 июля 2002 г.) / Под. ред. А.С. Нариньяни, - М.: Наука. 2002. Т.1. с.307-313.
19. Nastase, Vivi and Stan Szpakowicz. Word sense disambiguation in Roget's thesaurus using WordNet. In: *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.
20. Ng, Hwee Tou. Exemplar-based word sense disambiguation: some recent improvements. In: *Proceedings of the 2nd Conference on Empirical Methods in NLP (EMNLP-2)*, Providence, August 1997.
21. Ontology Usage and Application. // Applied Semantics. Technical Whitepapers. 2003.
22. Поляков В. Н., Бодров Д. А., Точин А. В. Интерактивные методы фокусировки и расширения поиска в интеллектуальной поисковой машине. Труды Международного семинара Диалог'2002. Протвино, 6–11 июня 2002 г. с. 438–449.
23. Поляков В. Н. Интеллектуальная поисковая машина. Концептуальный проект. Труды Казанской школы по компьютерной и когнитивной лингвистике. TEL-2000. Вып. 5. Казань. 17–20 октября. 2000 г. Казань. Изд-во Сэлэт. 2000. с. 87–119.
24. Richardson, R. and Alan F. Smeaton. *Using WordNet in a Knowledge-Based Approach to Information Retrieval.* Working paper CA-0395, School of Computer Applications, Dublin City University, Dublin, 1995.
25. Sussna, M. Word sense disambiguation for free-text indexing using a massive semantic network. In: Bhargava, B., T. Finin and Y. Yesha, eds., *Proceedings of the 2nd International Conference on Information and Knowledge Management*, Arlington, 1993, pp. 67 - 74.
26. Vorhees, Ellen M. Using WordNet for text retrieval. In: Fellbaum, Christiane, ed., *WordNet: An Electronic Lexical Database*, MIT Press, May 1998.
27. Wiebe, Janyce, J. Maples, L. Duan and Rebecca Bruce. Experience in WordNet sense tagging in the Wall Street Journal. In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?* Washington, April 1997.
28. Zipf, G. K. 1945. *The Meaning-Frequency Relationship of Words*. Journal of General Psychology 33: 251-256.
29. OntoQuery project. http://www.ontoquery.dk