УДК 519.686.4

В.А. Гужов, Л.Э. Чалая, А.В. Чижевский

МЕТОД ПОИСКА ИНФОРМАЦИИ В СОЦИАЛЬНОЙ СЕТИ «УЧЕНЫЕ УКРАИНЫ» ПО ОНТОЛОГИЧЕСКОЙ МОДЕЛИ

Аннотация. В работе проведен анализ проблемы оперативного поиска данных в социальных сетях. Показана перспективность применения онтологических моделей для реализации семантического подхода к обработке запросов пользователей социальных сетей. Построена онтологическая модель социальной сети «Ученые Украины», предназначенной для обеспечения координации научной и педагогической деятельности отечественных ученых. Предложен алгоритм семантического поиска информации по разработанной онтологической модели.

Ключевые слова: социальная сеть, онтологическая модель, семантический поиск, байесовская классификация, релевантность

Введение

В последнее время получили развитие социальные сети различного функционального назначения в качестве эффективного механизма взаимодействия пользователей глобальной сети Интернет. Социальная сеть, дополненная поддерживающими сервисами, представляет собой интерактивный многопользовательский веб-сайт, контент которого наполняется не только разработчиками, но и самими участниками сети [1].

На кафедре искусственного интеллекта Харьковского национального университета радиоэлектроники разрабатывается компьютерная социальная сеть «Ученые Украины», предназначенная для обеспечения координации научной и педагогической деятельности отечественных ученых. Одним из наиболее важных условий эффективного функционирования этой сети является реализация процедуры оперативного поиска необходимой информации. В частности, предполагается организация поиска по персональным данным ученых, зарегистрированных в сети, по группам, создаваемым пользователями, по форуму и по материалам, выложенным на сайте (публи-

 ISSN 1562-9945

кациям, объявлениям и т.д.). Основной задачей, возникающей при работе с распределенными полнотекстовыми коллекциями документов, является задача поиска документов по их содержанию. При этом критерии поиска могут быть основаны на анализе текстов (например, отыскать в сети конкретного пользователя можно путем нахождения текстовых фрагментов, содержащих информацию (личные данные, научные интересы, специальность), наиболее релевантную введенному запросу) [2,3]. Однако ставшие традиционными средства контекстного поиска по вхождению слов в документ зачастую не обеспечивают адекватного выбора информации по запросу пользователя. Одним из возможных вариантов решения этой проблемы является семантический поиск, который позволяет находить ресурсы не только по заданным словам из запроса, но и по эквивалентным по смыслу терминам. Например, по запросу «искусственный интеллект» при поиске по публикациям машина поиска сайта должна понимать, что статьи, с такими ключевыми терминами как «инженерия знаний», «робототехника», «обработка естественных языков» и др., также относятся к тематике запроса, даже если там нет непосредственного использования словосочетания «искусственный интеллект». Для эффективного семантического поиска необходима информация о предметной области, свойственных ей понятиях и отношениях между ними, а также ограничениях, свойственных этим отношениям. Такую информацию принято называть онтологией. Онтологии включают доступные для компьютерной обработки определения основных понятий предметной области и связей между ними.

Представляется перспективным применение в социальных сетях методов семантического поиска с использованием предварительно разработанной онтологической модели.

Постановка задачи

Целью настоящей работы является исследование возможности повышения оперативности обработки запросов пользователей социальной сети «Ученые Украины» на основе методов семантического поиска. В связи с этим были поставлены следующие задачи:

– построить онтологическую модель сети «Ученые Украины»;

– разработать алгоритм семантического поиска информации с использованием онтологической модели.

Онтологическая модель социальной сети

Для реализации на сайте семантического поиска была построена онтологическая модель социальной сети «Ученые Украины». Эта модель включает в себя следующие основные понятия: ученый; научная специальность; группа; публикация; сайт; новости; форум; сообщения.

Формально можно определить онтологию как множество

$$O = (L, C, F_l, F_c, R_h), \quad (1)$$

где $L = \{(w_i, x_i)\}_{i=1,n}$ – словарь терминов предметной области, w_i – термин; x_i – его рейтинг относительно других терминов концепции.

$C = \{c_i\} i = 1, m$ – набор понятий (концепций),

$F_l(L) \to C$ – функция интерпретации терминов – сопоставляет набору терминов из словаря подмножество концепций;

$F_c(C_i) \to L$ – функция интерпретации концепций – сопоставляет концепции набор терминов из словаря;

R_h – отношения иерархии между концепциями.

В качестве функции интерпретации терминов примем $P(c_i|u)$ – вероятность выбора концепции при условии запроса u.

Расширим словарь L, дополнив его определениями всех терминов и исключив вместе с тем понятие «рейтинг термина». Тогда L можно формально определить так: $L = \{v_i\}$, где v_i - это тексты, вмещающие в себя термин, его синонимы и определение.

Примем два предположения относительно слов в запросе u:

– все слова в запросе u являются одинаково важными;

– все слова в запросе u являются статистически независимыми, т.е. значение одного слова запроса однозначно не связано со значением других слов запроса.

Тогда функцию интерпретации $F_l(L) = P(c_i|u)$, используя формулы наивной классификации, можно представить следующим образом:

$$P(c_i|u) = \arg\max\left(\prod_i (P(w_j|c_i))\right), \quad (2)$$

где w_j – слова из запроса u,

$P(w_j|c_i)$ – вероятность принадлежности слова w_j к концепции c_i.

Очевидно, что:

 ISSN 1562-9945

$$P(w_j|c_i) = \frac{n_k + 1}{n + |L|}, \quad (3)$$

где n_k – количество упоминаний слова w_j в текущей концепции c_i,

n – общее количество различных слов в текущей концепции c_i,

$|L|$ – мощность множества слов во всех концепциях.

В итоге, применяя к исходному запросу функцию интерпретации (2), можно получить номер той концепции из формируемой онтологии (1), которая наиболее релевантна теме запроса. После чего необходимо расширить запрос терминами из найденной концепции с помощью функции интерпретации концепций $(F_c(C_i) \to L)$.

Поскольку в основе онтологии лежит древовидная структура концепций, то определив наиболее релевантную концепцию, получаем также список подчиненных концепций, термины которых могут служить дополнением к исходному запросу.

Организация семантического поиска данных по онтологической модели

Следует отметить, что не всегда результат поиска, содержащий одну концепцию, наиболее релевантную входящему запросу, дает возможность предоставить пользователю в качестве ответа действительно те данные, которые он хотел найти. Например, при введении запроса «А. Иванченко. Доказательство алгоритмической неразрешимости проблемы остановки машины Тьюринга» (поиск по публикациям) будет найдена концепция, соответствующая тематике запроса, а, следовательно, будет сформирован список всех зарегистрированных в сети публикаций по данной теме, в котором пользователь вынужден будет самостоятельно искать конкретную публикацию А. Иванченко.

Общая схема предлагаемого алгоритма семантического поиска информации с использованием онтологической модели социальной сети, позволяющего существенно снизить возможность возникновения таких ситуаций, приведена на рис. 1.

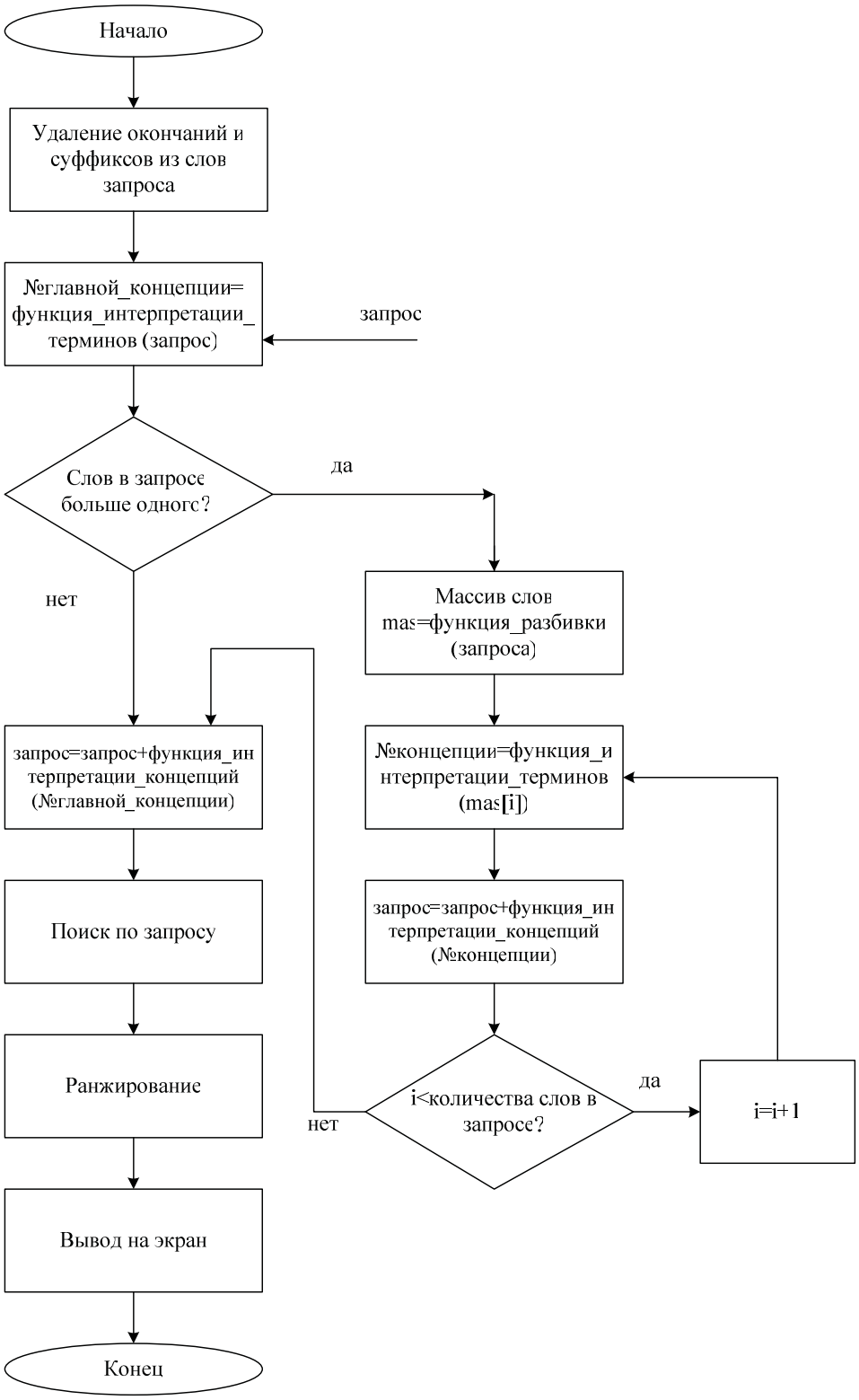

Рисунок 1 - Алгоритм поиска данных по онтологической модели

Его реализация предполагает выполнение поисковой системой социальной сети следующих действий:

 ISSN 1562-9945

– после того, как с помощью функции интерпретации найдена основная концепция, запрос разбивается на подзапросы, каждый которых содержит лишь одно слово исходного запроса;

– для всех сформированных подзапросов с помощью функции интерпретации вид (3) определяется номер концепции, после чего исходный запрос расширяется терминами (однако в него не добавляются термины из подчиненных концепций).

Такое расширение исходного запроса позволяет свести к минимуму возможность потери той информации, которая может быть важна пользователю, но при этом возникает проблема избыточности получаемых данных.

Частично избавиться от такой избыточности можно путем автоматического ранжирования получаемой по запросу информации по ее релевантности. Большинство запросов пользователей сайта социальной сети связано с поиском текстовых данных. В частности, задача поиска данных о конкретном ученом сводится к поиску текстовых фрагментов, содержащих информацию (личные данные ученого, его научные интересы, специальность и т.д.), наиболее релевантную введенному запросу. После расширения запроса (в соответствии с рассмотренным выше алгоритмом семантического поиска) формируется выборка ответов. Для того чтобы избавиться от избыточных ссылок, содержащихся в этой выборке, необходимо ранжировать полученные результаты. В результате такого ранжирования пользователь вначале получит наиболее релевантные ответы, а если его не удовлетворят данные, содержащиеся в первых ссылках, он может продолжить просмотр результатов поиска по расширенному запросу.

Ранжирование результатов может быть, например, осуществлено с помощью алгоритма байесовской классификации. Этот алгоритм позволяет оценить вероятность того, что текущий ответ (в виде текста) интересует пользователя, инициировавшего соответствующий запрос. При этом результатом поиска является формирование списка ответов, ранжированных по таким вероятностям.

В социальной сети «Ученые Украины» предполагается для поиска информации использовать также комбинированный подход, основанный, в частности, на семантическом поиске и поиске по ключевым словам (здесь в виде фрейма рассматривается HTML-страницы, а содержимое фреймов (слоты) – ссылки на отдельные страницы).

Результаты тестового моделирования подтверждают работоспособность рассмотренного подхода.

Выводы

Реализация эффективного семантического поиска в социальных сетях может быть осуществлена с использованием предварительно разработанной многокомпонентной онтологии, позволяющей устранять избыточность данных, получаемых пользователем сети в соответствии с запросом по заданной тематической категории. Предложенный в настоящей работе алгоритм положен в основу разработки социальной сети «Ученые Украины», предназначенной для обеспечения координации научной и педагогической деятельности отечественных ученых. Перспективным продолжением выполненных исследований является возможность дополнения схемы семантического поиска поиском по ключевым словам, используемых во фреймовой модели репозитория научных публикаций, зарегистрированных в базе данных социальной сети.

ЛИТЕРАТУРА

1. Gruber, T. R. A translation approach to portable ontology specifications / T. R. Gruber // Knowledge Acquisition. 1993. № 5(2).

2. Zakharova, I. V. An approach to automated ontology building in text analysis problems / I. V. Zakharova, A. V. Melnikov, J. A. Vokhmitsev // Workshop on Computer Science and Information Technologies, 2006. P. 177–178.

3. Melnikov, A. V. Method of automatic ontology creation based on bibliographic databases / A. V. Melnikov, I. V. Zakharova //Workshop on Computer Science and Information Technologies, 2005. P. 270–272.

 ISSN 1562-9945