УДК 004.03, 004.04, 004.09

АРХИТЕКТУРА РАСПРЕДЕЛЕННОЙ ИНТЕЛЛЕКТУАЛЬНОЙ ПОИСКОВОЙ СИСТЕМЫ, ИСПОЛЬЗУЮЩЕЙ ФАЙЛОВЫЕ НЕЙРО-ИНДЕКСЫ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ И НЕДОСТАТОЧНОСТИ АПРИОРНОЙ ИНФОРМАЦИИ О ДИНАМИКЕ РАЗВИТИЯ[1]

***Гарнага В. В.*[2], *Кольцов Ю. В.*[3], *Полупанов А. А.*[4], *Продан Ю. М.*[5]**

ARCHITECTURE OF DISTRIBUTED INTELLIGENT SEARCH SYSTEM USING FILE NEURO-INDEXES IN THE CONDITIONS OF UNCERTAINTY AND THE LACK OF A PRIORI INFORMATION ABOUT THE DYNAMICS OF THE SYSTEM

Garnaga V. V., Koltsov Y. V., Polupanov A. A., Prodan Y. M.

The article describes the architecture of distributed intelligent search system (DISS) and methods of presentation search indexes in the form of neural network weights. The proposed system allows you to track the dynamics of the DISS in the conditions of uncertainty and the lack of a priori information.

Keywords: uncertainty, neural network, search, index.

В настоящеее время информация является одним из самых мощных ресурсов, которыми располагает организация. Известно, что способность быстро и эффективно находить информацию увеличивает размер прибыли и является важным конкурентным преимуществом. По качеству построения информационной инфраструктуры предприятия можно судить об эффективности использования информации предприятием. Высокотехнологичная система, позволяющая быстро находить данные, необходимые для развития бизнеса, является признаком успешного предприятия. Для активно развивающихся средних и малых предприятий (SMB-сектор) [1] это является особенно актуальным.

Для создания эффективного алгоритма интеллектуального поиска информации проведено исследование в сферах интеллектуальных технологий и распределенных сред вычислений, а также разработано программное обеспечение, как практическая реализация результатов данных исследований. Указанная реализация ориентирована на применение в корпоративных сетях распределенной обработки данных и позволяет сократить расходы на информационные технологии благодаря использованию принципиально нового подхода к интеллектуальному информационному поиску.

Разработанная распределенная интеллектуальная поисковая система (РИПС) основана на общепринятых принципах и подходах, развивающих кластерную архитектуру вычислений [2]. Вычисления в кластерной архитектуре хорошо зарекомендовали себя во многих проектах. Последние достижения существенно упрощают создание программного обеспечения в рамках этой технологии. Организации, осуществляющие перевод массовых вычислений на модель распределенных вычислений, добиваются сокращения затрат на комплексы серверов приложений и

[1]Работа выполнена при поддержке РФФИ (13-01-00807).

[2]Гарнага Валерий Владимирович, канд. физ.-мат. наук, доцент кафедры информационных технологий Кубанского государственного университета; e-mail: Valeriy.Garnaga@gmail.com.

[3]Кольцов Юрий Владимирович, канд. физ.-мат. наук, заведующий кафедрой информационных технологий Кубанского государственного университета; e-mail: yurikoltsov@mail.ru.

[4]Полупанов Алексей Александрович, канд. техн. наук, доцент кафедры информационных технологий Кубанского государственного университета; e-mail: polualex@mail.ru.

[5]Продан Юрий Михайлович, преподаватель кафедры информационных технологий Кубанского государственного университета; e-mail: y.prodan@gmail.com.

административных расходов, а также отмечают улучшение управляемости своих систем. Кроме того, совокупная стоимость владения (Total cost of ownership, TCO) может существенно снизиться [3].

При работе с современными объемами данных (пета- и экзабайты) критически важными являются скорость и точность извлечения полезной информации. Для повышения эффективности получения данных, соответствующих критерию запроса пользователя, актуальной является проблема создания методологии интеллектуального поиска с использованием ассоциативной и семантической информации на основе искусственных нейронных сетей (ИНС).

Следует отметить, что результаты работы описываемой ниже технологии качественно отличаются от результатов, получаемых при применении алгоритмов нечеткого поиска, например основанных на использовании расстояния Левенштейна [4].

Для описания предлагаемой методологии рассмотрим процесс обучения ИНС (рис. 1), создающейся для каждого файла в отдельности. В этом случае индексной информацией для каждого файла являются веса этой ИНС. На вход ИНС подается один термин из словаря и порядковый номер вхождения этого термина в файл. Для упрощения структуры нейронной сети, представим длину всех терминов одинаковой и равной максимальной длине терминов. Все более короткие термины дополняются спецсимволами (например пробелом) в конце для достижения максимальной длины. На выходе нейронной сети формируется позиция вхождения термина в файл. Если вхождения с заданным порядковым номером не существует, то ИНС возвращает несуществующее значение (например, -1). Результатом работы ИНС является также количество вхождений термина, минимальная и максимальная позиции термина в файле.

Представим обучающее множество в виде множества $\mathbf{M} = \{(\mathbf{U}, \mathbf{V})\}$, где $\mathbf{U} = (a, b)$ — вектор, представляющий значения входов нейронной сети. Здесь a — последовательность символов, составляющих термин; b — номер вхождения термина, а в файл; $\mathbf{V}$ — вектор, представляющий значения выходов нейронной сети.

На основе обучающего множества M строится семантическая сеть (СС), что позволяет использовать предлагаемую методику для поиска информации на разных языках [4]. ИНС для разных языков для одного и того же файла могут отличаться, но результаты поисковых запросов по этим сетям совпадают, так как искомая фраза посредством СС преобразуется в некоторую универсальную форму. Некоторая потеря грамматической информации при этом не повлияет кардинально на результат поиска.

Отметим, что для обучения по предлагаемому методу необходимо временно сформировать классический индекс файла. На его основе и основе словаря терминов реализуется алгоритм обучения ИНС с обратным распространением ошибки.

Тип нейронной сети — сеть прямого распространения или персептрон. Из-за большого объема обрабатываемой информации используются скрытые слои. Выбор количественных показателей нейронной сети может осуществляться различными способами. В частности, это могут быть эвристики, генетические алгоритмы, экспериментальный подбор параметров.

В качестве входных значений нейронной сети помимо терминов существует возможность использовать дополнительные параметры. Например, параметр точности совпадения слов; «степень интеллектуальности»; параметр, характеризующий отношения между терминами на основе СС и т.д.

В качестве начальной ИНС для обучения на любом файле, предлагается взять ИНС, инициализированную такими весами, которые позволяют различать термины независимо от их формы. Для этого в самом начале работы ИПС специально подготавливается начальная ИНС. Она обучается на множестве, состоящем из всех известных терминов, но при обучении термины изменяются согласно грамматическим правилам. По окончании обучения (индексации) на файле из файловой системы проверяется корректность работы ИНС с грамматикой. Это является еще одним этапом тестирования результирующей ИНС для файла.

Для повышения эффективности работы РИПС предлагается использовать так называемый «any-time алгоритм», суть которого заключается в постепенном улучшении частных результатов. Т.е. известно, что наибольшее время обучения ИНС занимают последние итерации перед достижением требуемой точности. Если мы уменьшим требуемую точность на начальном этапе обучения на

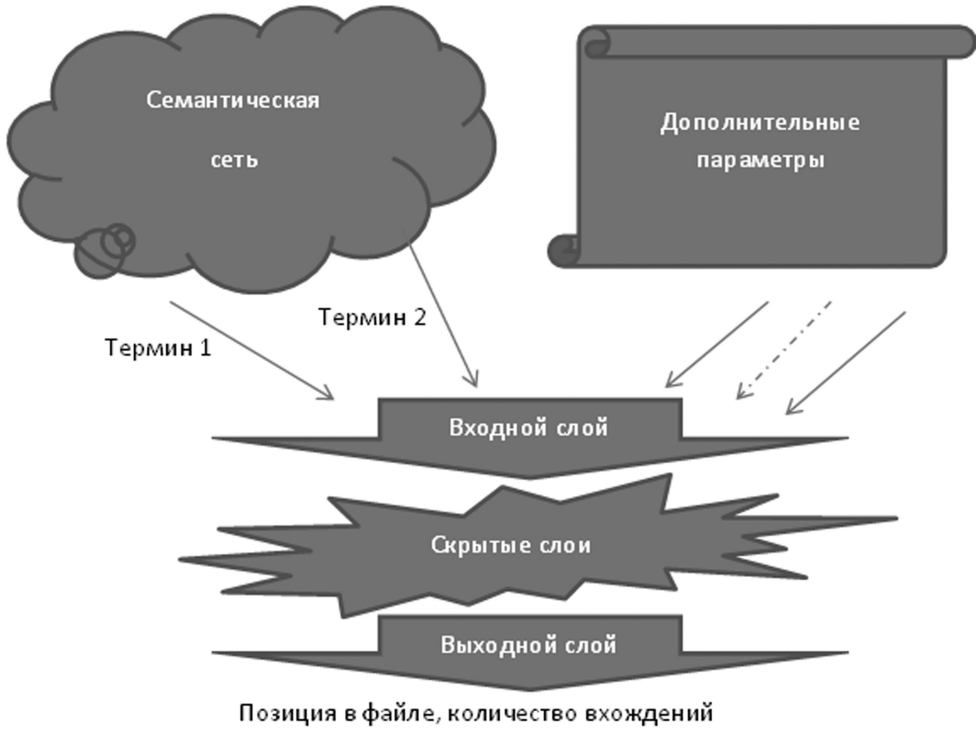

Рис. 1. Обучение ИНС

10 %, то время обучения может сократиться на 90 %. В дальнейшем, во время простоя системы, ИНС всегда возможно доучить. Как показывает практика, такой подход позволяет улучшить общий результат работы системы на 40–50 % [5].

Полученные в результате веса ИНС являются индексной информацией для каждого файла. Для ускорения процесса поиска в индексной информации используется ИНС Кохонена векторного квантования, обучаемая с учителем [6] (рис. 2). Для этого применяются следующие шаги:

1. Выбор очередного наблюдения осуществляется после индексации соответствующего файла. Вектор, представляющий входы и выходы ИНС, будет этим наблюдением.

2. Нахождение для него лучшей позиции соответствия узла на карте Кохонена. Вектор, веса которого меньше всего отличается от наблюдения в метрике, основанной на минимальном расстоянии между терминами, частотами встреч и синонимов этих терминов.

3. Определение количества соседей и обучение — изменение векторов веса и его соседей с целью их приближения к наблюдению.

4. Определение ошибки карты.

5. Переобучение по алгоритму DLVQ Fundamentals [7].

Опишем процесс работы ИНС:

1. Поисковая фраза разбивается на пары терминов. Например, фраза «что купить на праздник» разбивается на пары (что, купить), (купить, на), (на, праздник), (что, на), (что, праздник), (купить, праздник).

2. Последовательность поиска пар определяется в соответствии с порядком в поисковом запросе (что, купить), (купить, на), (на, праздник).

3. Используя ИНС Кохонена ищется множество индексов ИНС, в которые входят все эти термы.

4. Ранжируются файлы по результатам работы ИНС, полученных на основе весов из индексной информации с парами.

Очевидно, что для работы ИНС требуется много вычислительных ресурсов. Этот вопрос решается путем использования сред параллельных и распределенных вычислений. Существует множество работ посвященных использованию ИНС в параллельных и распределенных вычислительных средах [8, 9]. Применение развитых методов параллельных вычислений, которые позволяют добиться увеличения производительности на многоядерных процессорах, и методов распреде-

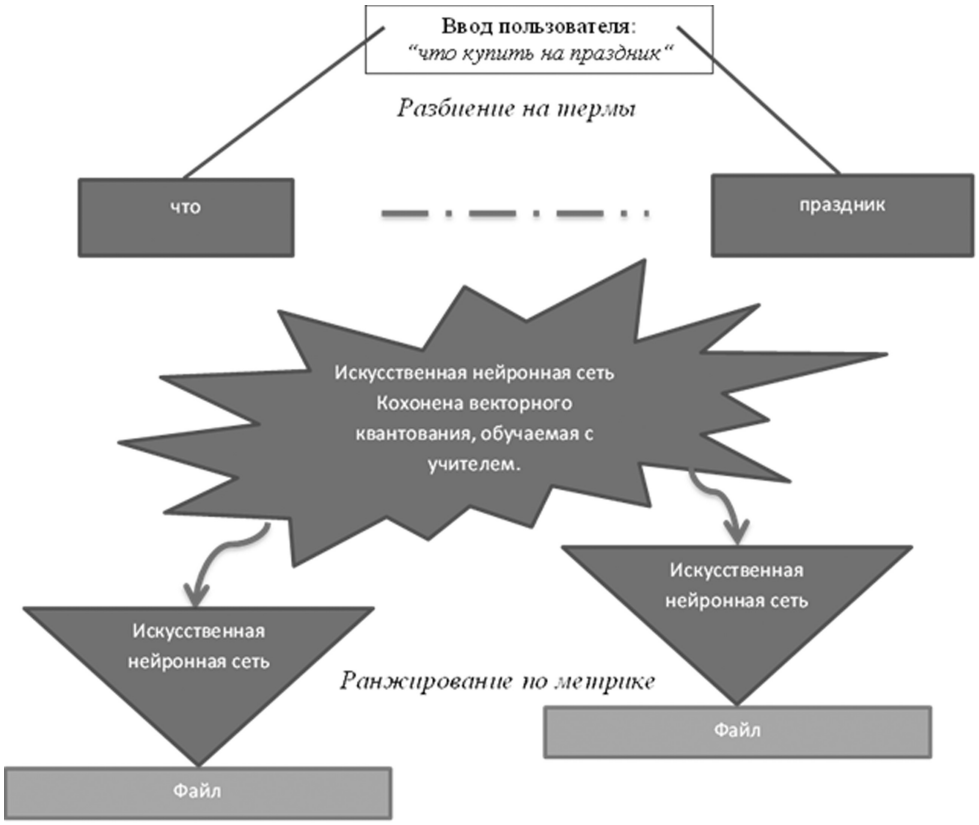

Рис. 2. Процесс работы РИПС

ленных вычислений, таких как MapReduce, дают возможность обрабатывать значительные объемы данных в РИПС.

Для функционирования РИПС необходимо обеспечить возможность автоматического динамического подключения и отключения дополнительных индексирующих сервисов. Система определяет для каждого интеллектуального индексирующего сервиса необходимую производительность, а также вычислительные элементы, которые этот сервис будет использовать.

Таким образом, рассматриваемая поисковая система имеет возможность использования множества компьютеров, на которые распространена инфраструктура распределенных вычислений и на которых выполняются индексирующие сервисы. Вычислительные мощности распределенной системы рассматриваются как единый пул ресурсов, динамически выделяемых при необходимости индексации того или иного файла в соответствии со стратегиями индексирования и предоставления ресурсов, а также с учетом состояния всей системы и ее компонентов. Для всех индексирующих сервисов существует набор стандартных функций. То есть все сервисы создаются на основе некоего абстрактного класса и имеют общие функции:

– функция инициализации работы сервиса настраивает параметры, необходимые для начала работы;

– функция корректного завершения работы сервиса закрывает соединения, освобождает память и выполняет другие необходимые функции;

– функция протоколирования работы сервиса записывает отчет о работе в файл. Данная функция очень полезна при анализе работы системы;

– функция, определяющая состояние или статус сервиса необходима при реализации протокола взаимодействия с сервисом;

– функция, идентифицирующая сервис и его параметры, необходима для последующего построения правильной структуры вызовов к этому сервису.

РИПС позволяет динамически подключать или отключать новые индексирующие сервисы. Администратор описывает правила распространения индексирующих процессов на дополнительных клиентов. Например, вычисления производятся на двух клиентах. РИПС постоянно измеряет нагрузку на уз-

лы и, если она превысит заданный в правилах предел, то на одном из разрешенных доступных клиентов автоматически запустятся дополнительные индексирующие процессы. Тем самым вычислительный ресурс РИПС увеличится. При дальнейшем увеличении нагрузки могут запускаться новые экземпляры индексирующих процессов. При снижении нагрузки клиенты будут освобождаться. Вычислительная нагрузка постоянно измеряется, вновь возникающие задачи индексации направляются на наименее загруженные клиенты. Тем самым достигается балансировка загрузки в соответствии со стратегией индексации. Администратор имеет возможность управления всеми сервисами (запускает, останавливает, конфигурирует множество клиентов), подключает новые компьютеры к РИПС. Можно создать несколько вариантов списков узлов и стратегий индексации и активизировать разные варианты в разные периоды времени.

РИПС реализует модель объединения вычислительных ресурсов организации в единую инфраструктуру, способную динамически настраиваться как в соответствии с изменяющимися требованиями бизнеса, так и по техническим причинам. РИПС делает возможным использование множества непроизводительных компьютеров для построения высокопроизводительной инфраструктуры информационного поиска, легко расширяемой в случае необходимости путём добавления новых вычислительных устройств. Кроме того, вычислительные ресурсы компьютеров могут динамически перераспределяться между различными вычислительными процессами, что повышает их полезную загрузку, с очевидной экономией расходов организации на приобретение дополнительного оборудования.

Литература

1. *Борисов С.* Конкурентоспособность и малое предпринимательство // Вопросы экономики. 2005. №1. С. 65–70.
2. *Сундеев П.В.* Разработка научно-методического аппарата анализа функциональной стабильности критичных информационных систем: Дис. ... канд. тех. наук, Краснодар, 2007.
3. *Zilberstein S.* Using anytime algorithms in intelligent systems // In Proc. of the Eighteenth International Joint Conference on Artificial Intelligence, 1996. P. 528–544.
4. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // ДАН СССР. 1965. Т. 163. Вып. 4. С. 845–848.
5. *Каяшева Г.А.* Дискретно-логические регуляторы с минимизацией продолжительности отработки системы продукционных правил и повышенной точностью: Дисс.... канд. техн. наук. Уфа, 2009. 153 с.
6. *Уоссермен Ф.* Нейрокомпьютерная техника: Теория и практика. М.: Мир, 1992. 240 с.
7. DLVQ Fundamentals. [URL: http://www.ra.cs.uni-tuebingen.de/SNNS/UserManual/node160.html].
8. *Aberdeen D., Baxter J., Edwards R.* 92c/Mflops, Ultra-large-scale neural-network training on a PIII cluster // In: Proceedings of Supercomputing 2000.
9. *Гарнага В.В.* Распределенная система прогнозирования FOREGRID // Студенческая научная весна–2009: Матер. Межрегиональной научно-технической конференции студентов, аспирантов и молодых ученых Южного федерального округа. Новочеркасск, 2009. С. 20–21.

Ключевые слова: неопределенность, нейронная сеть, поиск, индекс.

Статья поступила 14 июня 2013 г.
Кубанский государственный университет, г. Краснодар