Архитектура системы охвата информационных связей объектов мониторинга

Designing a system of managing information links between monitored objects

Ландэ Д. В. (dwl@visti.net), **Брайчевский С. М.** (smb@visti.net), **Дармохвал А. Т.** (hval@visti.net), **Жигало В. В.** (vladlen@visti.net)

Информационный центр «ЭЛВИСТИ», Киев, Украина

Представлен подход к построению полнотекстовой информационно-поисковой системы, основными элементами которой являются не отдельные термы, а взаимосвязи между понятиями, экстрагируемыми из текстовых документов. Описаны основные компоненты и архитектурные решения, применяемые в данной системе, а также интерфейс пользователя.

В настоящее время информационное пространство представляет собой динамическую среду, наполнение которой постоянно изменяется. Существующие доступные фактографические базы данных структурированной информации не всегда могут прийти на помощь исследователю-аналитику. Для оперативного определения фактов и сущностей, моделирования информационных связей между ними наиболее перспективным подходом оказывается учет информации, знаний, которые содержатся в неструктурированных текстовых документах, в частности, в Интернет. Поиск в базах данных неструктурированной текстовой информации может применяться для задач наведения исследователей-аналитиков «на цель» в условиях, когда фактографические базы данных структурированной информации труднодоступны, неполны, неоперативны.

Неструктурированные тексты содержат в себе несравненно больше важной информации, чем структурированные записи баз данных, именно в силу того, что формализации подлежит сравнительно небольшой сегмент информации. В настоящее время появляется все больше качественных инструментальных средств извлечения понятий из неструктурированных текстов [1], [2], [3].

Сегодня ни у кого не вызывает сомнений то, что не удастся создать единую универсальную информационно-поисковую систему, которая решала бы одинаково эффективно все поисковые задачи. Поэтому представляется актуальным поиск технологических решений, ориентированных на применение к различным предметным областям, предполагающим определенные информационные потребности.

Ниже будет представлено одно из решений, основанное на использовании в процессе полнотекстового информационного поиска взаимосвязей между понятиями.

В настоящее время, когда у пользователей уже накоплен большой опыт работы с традиционными информационно-поисковыми системами, оказалось очевидным, что факты или понятия, которые ищутся с помощью таких систем, сами по себе зачастую бессмысленны. Например, если пользователя интересуют информационные связи Сбербанка России с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а все документы, содержащие словосочетание «Сбербанк России» физически невозможно. В таких случаях информационные связи, интенсивность которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. «..Важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. Именно взаимосвязь способствует пониманию мотивационно-целевых особенностей отношений человека...» [4] . То есть пользователя интересует не понятие само по себе, а понятие в окружении, чтобы сразу иметь представление о предметной области, при необходимости направить уточняющий поиск в нужном направлении. Элементы такого подхода можно видеть, например, в «облаках» системы Quintura (http://quintura.ru), но там отображаются не понятия/сущности, а наиболее часто используемые термы.

Таким образом объективно существует необходимость построения эффективной полнотекстовой

информационно-поисковой системы, обеспечивающей поиск не по отдельным термам или понятиям, а по взаимосвязям между сущностями, присутствующими в документах, то есть создания систем, которые будем условно называть «базами данных связей» (БДС).

База данных практически любой традиционной информационно-поисковой системы может быть рассматриваться в виде графа, вершинами которого выступают объекты — термы, понятия, дескрипторы и др., а ребрами — их связи. Вместе с тем, основа поиска в этих случаях — поиск вершин, то есть поиск объектов. Поиск по взаимосвязям, ребрам, кажется на первый взгляд менее эффективным. Действительно, если предположить, что в графе N вершин, то ребер теоретически может составлять $N(N\text{-}1)/2$, то есть, если предположить, что вершин всего 100 тыс., то ребер может оказаться около 5 млрд, что соответствует достаточно большой базе данных даже по современным понятиям. Вместе с тем, если в качестве вершин графа использовать такие понятия, как имена людей и названия компаний из новостных документов, то оказывается, что соответствующая матрица инцидентности оказывается очень разреженной. Измерения показали, что при количестве отдельных понятий, извлеченных из 5 млн. новостных документов, равном примерно N = 1,5 млн., количество связей составило всего лишь v = 4 млн., то есть коэффициент разреженности матрицы данного графа составил:

$$K = \sqrt{\frac{v}{N(N-1)/2}} \approx 0,002.$$

Кроме того, как показали эксперименты, распределение степени вершин в подобных графах — степенное (см. рис. 1) [5], что свидетельствует о так называемой безмасштабности, то есть о том, что многие характеристики (в частности, соотношение количества вершин и ребер), должно оставаться на одном уровне. Поэтому в качестве основы построения базы данных связей сегодня оказывается технически возможным использование ребер рассматриваемого графа — связей между отдельными понятиями.

Исходя из результатов исследований, была создана база данных связей объектов путем мониторинга интернет-ресурсов. Пользователи этой системы фактически получают доступ к базе данных информационных взаимосвязей интересующих их объектов. Под информационной взаимосвязью в узких рамках представленной ниже системы понимается совместное упоминание объектов в некоторой информационной единице, принятой в качестве основной. Такими единицами могут быть документы, разделы документов, абзацы и т. д. Описываемая ниже реализация охватывает обычные статистические связи, однако, следует отметить, что в рамках предлагаемого подхода взаимосвязи могут быть определены и другими способами; в этой работе не идет речь о создании средств выделения понятий из текстов и установления связей между ними. Предлагается подход к созданию ИПС, в которой подобные средства играют важную, но вспомогательную роль. То есть в рамках данной работы речь не идет о создании высокоэффективной системы выявления связей, которая обладала бы явными преимуществами по сравнению с другими подобными системами. Вместе с тем речь идет о поисковой системе, ориентированной на использование конечным пользователем в режиме онлайн.

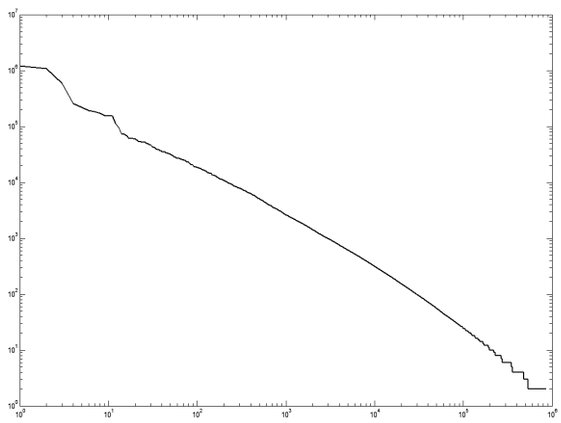

Рис. 1. Ранжированный график степеней вершин графа понятий (количества исходящих ребер) в логарифмической шкале

В качестве массивов новостной информации использовались фрагменты базы данных системы контент-мониторинга InfoStream [6], а также результаты мониторинга специализированных вебслужб, таких как базы данных биографий людей, организаций, служб трудоустройства и т. п.

Информационные взаимосвязи между понятиями выявляются путем обработки текстовых массивов и хранятся в специальной базе данных. Набор понятий, используемый при построении БДС, формируется путем экстрагирования данных из того же текстового массива, что придает системе целостность.

В корпоративной информационной инфраструктуре база данных связей может использоваться различным образом, например, отдельно, либо возможности БДС могут быть дополнены возможностями существующих полнотекстовых и/или фактографических баз данных (рис. 2). При этом основным результатом работы БДС является построение карт связей, а в качестве побочного эффекта, реализующего «режим доказательства», может рассматриваться извлечение самих документов как источников связей.

При проектировании БДС использовались решения, которые можно отнести к самым пер-

спективным в области создания информационно-аналитических систем, в частности, теория и технологии глубинного анализа тестов — Text Mining [7], в том числе развитая методология экстрагирования понятий [1, 2, 8], теория и технологии баз данных сверхбольших объемов, концепция «сложных сетей» (complex networks) [9, 10]. Теория сложных сетей изучает характеристики, учитывая не только на топологию сетей, но и статистические феномены, распределение весов отдельных вершин (в качестве которых можно рассматривать сущности, понятия, факты) и ребер, эффекты протекания и проводимости в сетях и т. п.

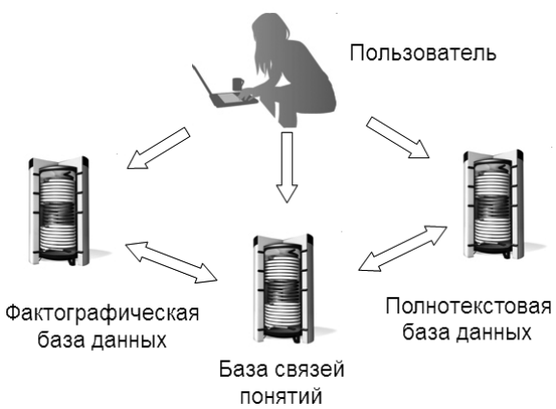

Рис. 2. Место базы данных связей понятий в корпоративной информационной инфраструктуре

На рис. 3 схематически представлены основные технологические этапы формирования базы данных связей. С помощью программы-робота осуществляется сканирование выбранных веб-ресурсов, которые содержат информацию, относящуюся к объектам исследований. После этого осуществляется экстрагирование необходимых пользователям понятий, например, имен персон, наименований брендов, компаний, электронных адресов и т. п. Отобранные понятия и соответствующие отношения между ними, загружаются в базу данных связей, которая также содержит ссылки на документы-первоисточники. Средства экстрагирования понятий ориентированы на обработку документов, сканируемых из Интернет, представленных как на русском, украинском, так и на английском языках. Реализовано автоматическое функционирование системы в режиме мониторинга интернет-ресурсов по мере их поступления.

Для настройки средств экстрагирования понятий предусмотрены таблицы шаблонов, такие как таблица фамилий известных персон, слов, заведомо не являющихся фамилиями (стоп-словарь), возможных имен, изменяемых окончаний и соответствующих им нормальных форм. Для экстрагирования названий компаний предусмотрены такие таблицы, как названия известных компаний, «префиксов», используемых для выявления неизвестных заранее компаний (применяется для русско- и украиноязычных документов), например, «АО», «ООО», «ТОВ», «АОЗТ» и др., «суффиксов», используемых для выявления неизвестных заранее компаний (применяется для англоязычных документов), например, «Ltd», «Inc», «Corp» и др.

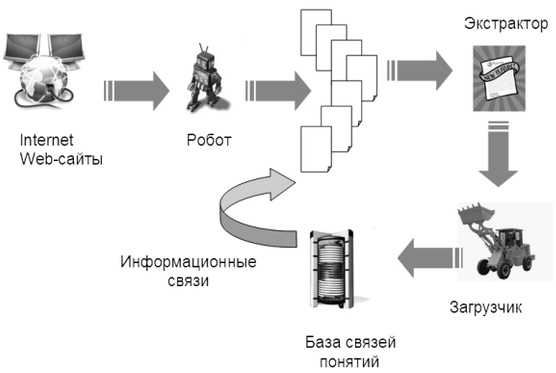

Рис. 3. Схема формирования базы данных связей

Предложенный подход к поиску, естественно, влечет за собой некоторые особенности в реализации архитектуры базы данных связей понятий. Кроме того, архитектура БДС должна быть ориентирована на такие возможные применения, как выявление неявных связей (не выявленных явно комплексом экстрагирования понятий), поиск отдельных объектов, а также взаимосвязь с существующими фактографическими базами данных. Вариант такой архитектуры в настоящее время реализован и используется в качестве компоненты системы конкурентной разведки X-Files.

Сама база данных связей состоит из двух разделов:

1. CTU (Concept-Time-Unq. number), отражающего появление понятия в текстах документов, основная таблица которого содержит такие поля, как понятие, дата, время упоминания, уникальный номер документа, абзаца, предложения.
2. CCN (Concept-Concept-Number), отражающего совместное появление пары понятий в текстах.

Информационные взаимосвязи объектов мониторинга расширяется за счет взаимодействия с фактографическими базами данных. Расширенная система фактически включает две сети. Первая сеть условно называется «Картой явных связей», а вторая, соответственно, «Картой вероятных связей» (рис. 4).

Для построения карты явных связей используется наложение двух промежуточных сетей (или слоев) — информационной и фактографической. Информационная сеть формируется из понятий,

экстрагируемых из текстовых документов, а фактографическая — на основании фактографических данных. Важная проблема связана с необходимостью выявления неочевидных закономерностей и связей понятий. На сегодня известно несколько путей решения этой проблемы, например, на основе концепции сложных сетей [11]. В настоящее время авторами проводятся исследовательские роботы по формированию карты вероятных связей, которые базируются на подходе, описанном в [12].

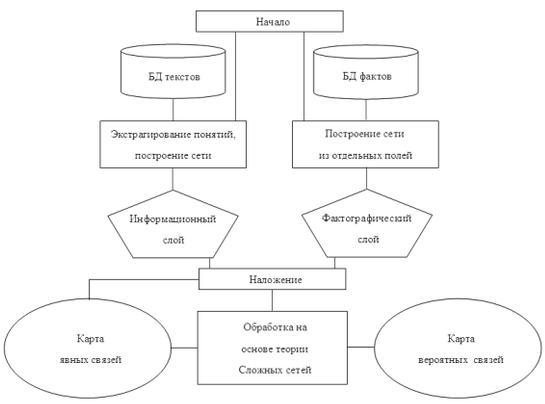

Рис. 4. Схема формирования карт явных и вероятных связей

Система позволяет пользователю в онлайн-режиме получать карты связей (КС) для указанных им объектов и помогает интерпретировать результаты. Предусматривается, что пользователь вводит в качестве запроса системе объект. Запрос направляется к БДС, откуда выбираются соответствующие ему фрагменты — карты связей (уровень детализации и временная ретроспектива должны указываться параметрически).

После выявления релевантных объектов и связей выполняются процедуры их автоматической группировки (кластеризации) и визуализации, результаты предъявляются пользователю в виде КС. Карты связей представляются в нескольких форматах, в том числе в табличном (таблица взаимосвязей понятий), графическом (круговая диаграмма), в виде динамических Java-диаграмм (графов связей), построенных с помощью средств TouchGraph.

Интерфейс взаимосвязей субъектов позволяет пользователю выбрать:

- вид субъекта;
- наименование субъекта;
- промежуток времени;
- глубину детализации.

Граф связей (ГС) строится с помощью апплетов Java и представляет собой графический объект, который содержит в своем составе узлы и ребра. Каждый элемент ГС имеет контекстное меню, которое является дополнительным элементом управления в интерфейсе пользователя БДС (рис. 5).

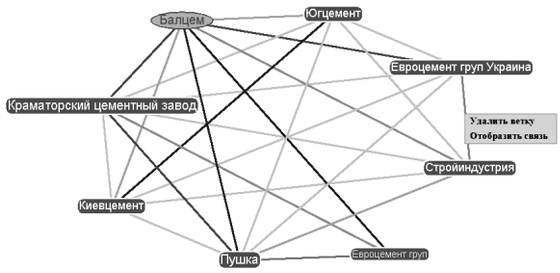

Рис. 5. Контекстное меню ребра

Объекты, которые имеют больше количество связей, изображаются с помощью большего шрифта. Ребра, соответствующие большему количеству связей, изображаются более темными линиями. Построенная сеть имеет собственные средства управления: изменение масштаба (с помощью меню «масштаб» или полосы прокрутки в верхней части экрана); перемещение всего графа; перемещение объекта; изменение конфигурации; подсветка связей выбранного узла и т. п.

В *таблице взаимосвязей понятий* данные представлены в виде квадратной таблицы, строки и столбики которой соответствуют объектам, связанным с исходным (рис. 6). Ячейки таблицы окрашивающиеся разными оттенками серого цвета, в зависимости от количества взаимосвязей объектов, которым соответствует строка и столбец. При наведении на ячейку таблицы указателя «мыши» на экран выводится количество взаимосвязей данной пары объектов.

Фирмы		1	2	3	4	5	6	7	8	9	10	11	12
Балцем	1												
Евроцемент груп-Украина	2												
Краматорский цементный завод Пушка	3												
Стройиндустрия	4												
Бахчисарайский комбинат Стройиндустрия	5												
Укрцемент	6												
Киевцемент	7												
Югцемент	8												
Dyckerhoff AG	9												
Волынь-Цемент	10												
Николаевцемент	11												
Альтцем	12												

Рис. 6. Таблица взаимосвязей понятий

Круговая диаграмма представляет собой графический объект, в котором узлы (все кроме главного равномерно распределены на кругу, а основной занимает центральное положение) соответствуют отобранным объектам, а ребра — связям между ними. На круговой диаграмме (рис. 7) ребра, соответствующие большему количеству связей, изображаются более толстыми и темными линиями. При нажатии на узел круговой диаграммы кнопкой «мыши» открывается список документов, в которых упоминается выбранный объект совместно с основным.

Приведем еще один пример использования БДС, случай, когда пользователя интересуют информационные связи Сбербанка России. Разумеется, для

запроса «Сбербанк России» может быть выявлено множество различных связей, но при этом существует простой и надежный критерий ранжирования результатов, состоящий в отсечении статистического фона. В рассматриваемом случае, задав соответствующий запрос можно получить граф (рис. 8) наиболее связанных со Сбербанком России объектов (персон и компаний). И если нахождение фамилий руководителей банка (председателя правления, первого заместителя председателя правления и руководителя дочернего банка) является достаточно очевидным результатом, то связи между отдельными банками позволили выявить (после обращения к документам-первоисточникам) неочевидные на первый взгляд факты, например, то, что УкрСиббанк и УкрСоцбанк являются банками-партнерами.

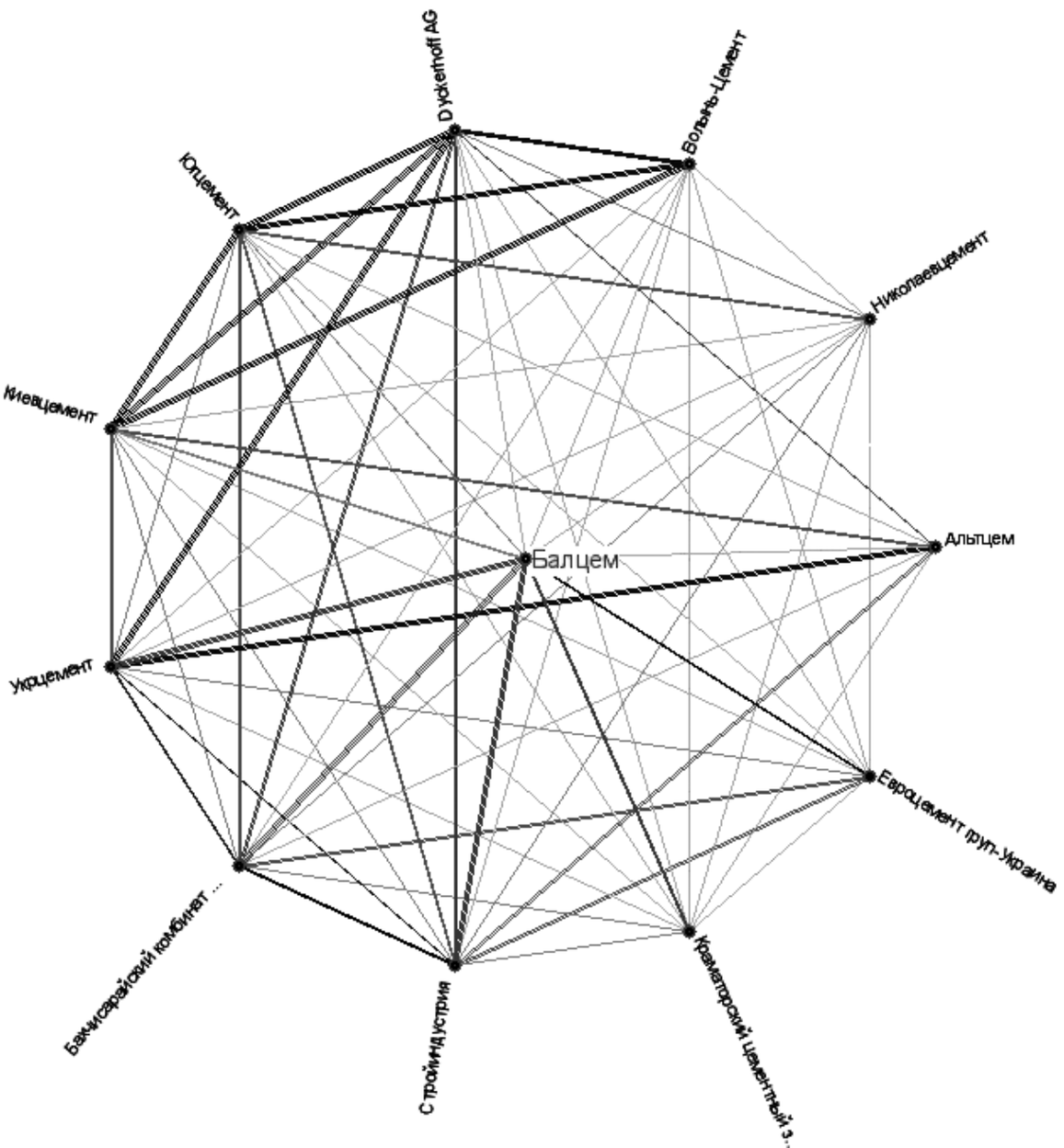

Рис. 7. Круговая диаграмма

Представленный подход может рассматриваться как основа построения информационно-поисковых систем, в которых изначально решены вопросы оперативности, отсеивания информационного шума. Рассматриваемая реализация — БДС имеет свойство масштабирования по трем параметрам: объему баз данных, составу понятий, которые используются, и по инфраструктурному окружению.

К настоящему времени еще не получено количественных оценочных характеристик для реализованного варианта системы, возможно при распространении представленного подхода, поиск по связям станет одной из дорожек TREC или РОМИП, как это происходило с подключаемыми к ней компонентами. Отдельные компоненты описанной системы, такие как модули экстрагирования и определения взаимосвязи понятий, средства визуализации и кластеризации, выявления «неявных связей», и т. п. могут заменяться и привязываться к информационно-поисковой системе с помощью прикладных программных интерфейсов.

Анализируя связи в сети, можно определить многие неочевидные свойства, например, выявить наличие кластеров, определить их состав, различия в связности внутри и между кластерами, идентифицировать ключевые элементы, которые связывают кластеры между собой и т. п. Серьезным препятствием при анализе является неполнота информации о связях между отдельными узлами сети. Вместе с тем сегодня уже существуют алгоритмы [11, 12], с помощью которых становится возможным с высокой вероятностью восстановить отсутствующие фрагменты связей. Даже не имея полного описания информационной сети, можно получать репрезентативную выборку «реальных» связей и по ней достроить всю сеть. Перспективы развития созданной системы — усложнение учитываемых связей, учет семантики контекста понятий в документах при их экстрагировании, отбор перечня действительно полезных баз данных текстовых документов, учет большего количества сущностей (понятий).

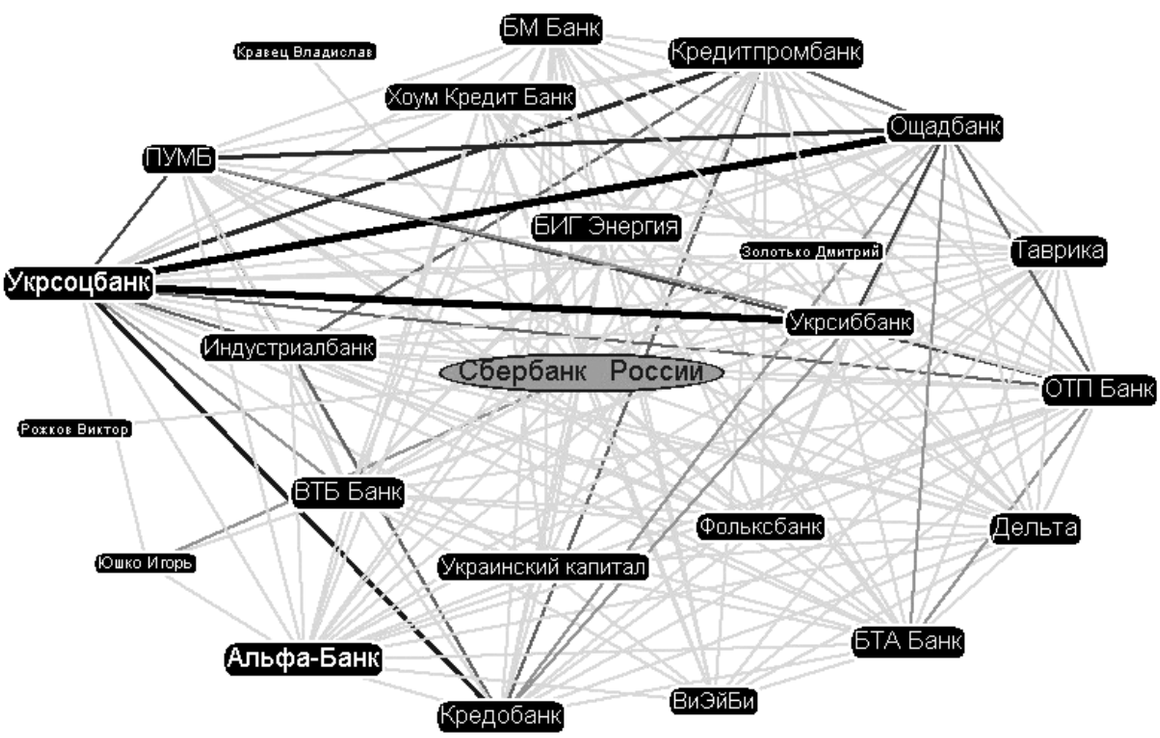

Рис. 8. Граф информационных связей понятия «Сбербанк России»

Представленная система является связующим звеном связи между полнотекстовыми и фактографическими базами данных. Очевидно, что реальный прорыв в области информационно-аналитической работы возможный лишь в результате агрегирования разных направлений. Базирующиеся на нескольких конкурирующих ранее точках зрения подходы на сегодня могут рассматриваться как пути создания современной мощной информационно-аналитической системы.

Литература

1. *Grishman R.* Information extraction: Techniques and challenges. In Information Extraction (International Summer School SCIE-97). Springer-Verlag, 1997.
2. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.)/ Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.: Наука.
3. *Додонов А. Г., Ландэ Д. В.* Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга // Регистрация, хранение и обработка данных, 2006, Т. 8, № 4. — С. 45–52.
4. *Массон Г. В.* Взаимосвязь системы личностных терминальных ценностей и типов межличностных отношений: Дис. ... канд. психол. наук : 19.00.01: Красноярск, 2004. — 146 с. РГБ ОД, 61:05–19/11
5. *Ландэ Д. В., Григорьев А. Н., Брайчевский С. М., Дармохвал А. Т., Снарский А. А.* Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2007, Переславль-Залесский, Россия, 2007. — С. 148–150.
6. *Григорьев А. Н., Ландэ Д. В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.)/ Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.:Наука — С. 109–111.
7. *Berry M. W.* Survey of Text Mining. Clustering, Classification, and Retrieval. — Springer-Verlag, 2004. — 244 p.
8. *Ермаков А. Е.* Автоматическое извлечение фактов из текстов досье: опыт установления // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 172–177.
9. *Newman M. E. J.* The structure and function of complex networks // SIAM Review. — 2003. — Vol. 45. — pp. 167–256.
10. *Ландэ Д. В., Снарский А. А., Безсуднов И. В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. — М.: Либроком (Editorial URSS), 2009. — 264 с.
11. *Clauset A., Moore C., Newman M. E. G.* Hierarchical structure and the prediction of missing links in networks // Nature. — 2000. — Vol 453. — pp. 98–101.
12. *Снарский А. А., Ландэ Д. В., Женировский М. И.* Метод выявления неявных связей объектов // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009, Петрозаводск, Россия, 2009. — С 46–49.