КОНЦЕПЦИЯ ОБЩЕСТВЕННЫХ БАНКОВ ДАННЫХ И ЗНАНИЙ

Куракин П.В., Институт проблем управления им. В. А. Трапезникова РАН
Митин Н.А., Институт прикладной математики им. М. В. Келдыша РАН

Аннотация. Развитие гражданского общества в России создает социальный заказ на информационные системы нового поколения. Основной задачей таких систем видится обеспечение возможности для граждан получать верифицируемую информацию о социально – экономическом положении страны. Наличие доступной верифицируемой информации представляется самым надежным залогом от манипуляции общественным сознанием со стороны деструктивных, антисоциальных и антигосударственных сил, где бы не находился их источник. Ниже мы предлагаем вариант возможной концепции такой информационной системы.

Мы рассматриваем предлагаемую информационную систему именно как инструмент гражданского общества, то есть – систему, доступ к которой и работа с которой рассчитан на «среднего человека». В настоящее время в нашей стране разрабатываются информационные системы, рассчитанные на экспертное сообщество, занятое прогнозированием социально-экономического развития России [1, 2]. Такие системы, безусловно, необходимы и сами по себе, и как возможные партнеры более демократичных информационных сетей. Например, одним из критериев хорошей функциональности предлагаемой нами системы может быть наличие экспорта данных в интересах других систем. Однако, исходно мы отталкиваемся от базовой идеи общественного банка данных как «народной социально-экономической информационной системы».

Мы исходим из следующих примерных критериев, которым должны отвечать общественные банки данных с точки зрения предполагаемого

пользователя – «обычного» гражданина, который заинтересован в обмене информацией социально-экономического характера о ситуации в России:

1. Система имеет дело, в общем случае, с самой разнородной (плохо структурированной) информацией.

2. Система должна быть открыта на чтение вообще для всех граждан России.

3. Крупномасштабная структура данных должна отличаться простотой и обозримостью, должен иметься простой и эффективный механизм поиска данных.

4. С точки зрения создания новых записей, система должна быть защищена от заведомо недобросовестной информации (спам, реклама, манипуляция и явная дезинформация).

5. Система поддерживает принцип «доступны все точки зрения, информация только добавляется, изменять и удалять запись может только ее автор».

6. Процесс верификации обеспечивается развитым инструментом комментирования и критики выложенных записей.

С технической точки зрения, требования к системе представляются следующими (частично перекрываются с «пользовательскими» критериями):

- простота архитектуры;
- простота обслуживания и использования;
- прозрачная дисциплина доступа по всей территории Российской Федерации;
- поддержка хранения слабо структурированной и слабо формализованной информации;

- поддержка эффективного поиска данных. Более подробно данные требования и пути к их исполнению обсуждаются далее в тексте.

Далее в тексте описаны предлагаемые нами принципы организации и функционирования банков и знаний, которые отвечают перечисленным базовым требованиям.

Принципы функционирования с точки зрения предполагаемого пользователя

Мы полагаем, что отправной точкой в проектировании описываемой системы должен быть тот факт, что она должна быть основана на слабо структурированных и плохо формализованных данных. Ниже будет описана предлагаемая структура единичной создаваемой записи. Крупномасштабная структура всей совокупности записей основана двух механизмах: ссылки и хештеги (плоская структура со связями и тематическими разделами, нет никакой иерархии). Процесс верификации информации обеспечивается механизмом специальных «критических» ссылок.

Лучше всего принять, что базовой единицей данных, размещаемых в системе, является простой текст, т.е. строка ASCII - символов. К тексту могут быть прикреплены любые дополнительные фрагменты данных, как-то: изображения (например, форматы bmp, jpeg, gif, png), таблицы Excel, текстовые файлы специализированных форматов (pdf, ps, doc, docx, html). Понятно, что содержательно, скорей всего, ASCII – текст будет только комментарием к прилагаемым документам – статьям, статистическим таблицам, графикам и пр. Однако, *технически* следует принять, что эти документы являются опциональными, а обязательным компонентом является именно ASCII – текст.

Иными словами, элементарная запись, хранящаяся в системе, состоит из нескольких компонентов (назовем их полями), главным из которых является поле простого текста. Это поле не только главное, но и обязательное. Также, у записи обязательно наличествует поле *автор*, значение которого создается автоматически – оно тождественно имени зарегистрированного пользователя системы, создавшего запись (о принципах регистрации пользователей см. далее).

Далее, представляется практически очевидным, что описываемая система имеет смысл только при наличии некоего инструмента, связывающего разные записи, хранящиеся в системе. Поэтому мы предлагаем следующее – не обязательное, но желательное – поле записи: *авторские ссылки*. Авторские ссылки представляют собой реализованные тем или иным техническим способом указатели на другие записи, хранящиеся в системе, на которые посчитал нужным указать автор записи. Важно, что это поле может редактировать только автор данной записи.

Далее, представляется разумным ввести поле *ссылки пользователей системы*. Это поле, в отличие от авторских ссылок, должно быть недоступно автору записи для редактирования. Смысл этого поля в том, что другие пользователи системы могут посчитать, что данная запись, по своему содержанию и своей логике, связана с записями, отличными от тех, которые считает связанными автор записи. Мы предлагаем реализовать такую идеологию системы (об этом подробнее будет сказано ниже), что принципиально должны присутствовать все возможные точки зрения. С одной стороны, есть тексты и ссылки, сообщенные их автором, с другой стороны - есть мнение сообщества об этих данных, и следует сохранять в открытом доступе и то, и другое. Для реализации этого принципа у каждой должны быть оба поля: авторские ссылки и ссылки пользователей системы. В первом приближении можно считать, что к

полю ссылок пользователей можно только *добавлять* значения (единичные ссылки), но нельзя удалять. Это тоже соответствует идеологии «присутствуют и доступны все точки зрения».

Представляется необходимым, чтобы пользователи имели возможность вносить в систему сообщения, представляющие собой *критику* (в том или ином смысле) других сообщений, созданных другими пользователями – это соответствует указанной идеологии «присутствуют и доступны все точки зрения». Мы полагаем, что сами такие сообщения не должны иметь специального статуса или быть сообщениями отдельного вида. Скорее, ссылки *на такие сообщения* (но, вероятно не *из таких сообщений*) должны быть отдельного вида. Мы предполагаем и предлагаем ввести еще одно поле в базовую структуру сообщения - *ссылки на критику*. В этом поле указывается список ссылок на сообщения, содержащих ту или иную критику данного сообщения. Задается ли значение такого поля автоматически или вручную – предстоит еще продумать. Важно то, что в самом критическом сообщении, видимо, нет необходимости указывать ссылку на критикуемое сообщение в отдельном поле – достаточно задать такую ссылку в поле авторских ссылок.

Далее, в системе должен быть организован механизм поиска информации. Представляется разумным, чтобы поиск был основан на тематических разделах, реализованных в виде *хештегов* [3], прикрепляемых к сообщению) - аналогично тому, как это делается в блоговых системах типа livejournal.com, blogspot.com, а также в социальной сети facebook.com и т.п. В системе необходимо запланировать разработку языка запросов на основе логических операций объединения, пересечения и вычитания множеств («найти сообщения с тегами A, B, C, но не включающие сообщения с тегами D и E»). Разумеется, можно обсуждать и другие критерии для поисковых запросов.

Опять же, имеет смысл иметь в сообщении два различных поля: поле авторских хештегов (которое может редактировать только сам автор) и поле хештегов других пользователей системы (по аналогии с полями авторских ссылок и ссылок других пользователей). Смысл этого в том, что автор имеет право монопольно задать для собственной записи только те тематические хештеги, которые считает нужным; в свою очередь, авторская оценка тематической принадлежности его сообщения вполне может не разделяться другими пользователями; они также имеют право высказать свое мнение о тематической принадлежности данной записи, не затрагивая авторской оценки. Как и в случае с полем ссылок пользователей, имеет смысл разрешить только добавление новых хештегов в список, но не удаление.

Таким образом, предварительно структура базовой единичной записи в системе выглядит так:

- уникальный (и глобальный, т.е. в пределах всей системы) идентификатор записи;
- автор записи;
- комментарий (ASCII – строка, основное и обязательное поле);
- блок прикрепленных данных (тексты, изображения, таблицы – необязательное поле);
- авторские ссылки (список ссылок на другие единичные записи);
- ссылки других пользователей;
- ссылки на критику данной записи;
- список авторских хештегов;

- список хештегов других пользователей системы.

Общие принципы хранения, обработки и использования данных

Исходя из вышесказанного, может возникнуть ощущение, что предлагаемая система напоминает всемирную общественную энциклопедию «Википедия». На самом деле, совпадение есть только в том, что любой желающий (лучше зарегистрированный пользователь) может вносить знания. Далее следуют фундаментальные отличия. Общие принципы хранения, обработки и использования данных в предлагаемой системе удобно объяснить, отчасти отталкиваясь от сравнений с системой «Википедия».

1. Согласно правилам системы «Википедия» [4], вносимые пользователями сведения должны основываться на уже опубликованных учебниках, монографиях, справочниках, монографиях. Общественный банк данных и знаний должен приветствовать наличие ссылок на руководства подобного рода, но главный его смысл – опубликование и обеспечение глобального доступа к информации, изначально доступной только локально (в территориально, или производственно – отраслевом смысле).

2. Формально, любой пользователь Интернета, даже не будучи зарегистрированным пользователем системы Википедия, может редактировать любую статью. Однако, существует общественный наблюдательный совет, который следит за тем, чтобы не было «хулиганской порчи» статей. На деле, особенно в части статей по общественно-политической тематике, это выливается в политическую борьбу разных «партий» в этом наблюдательном совете – например, из статей по политической истории России советского времени могут произвольно исчезать целые куски текста. Концепция общественного банка знаний прямо противоположна: любое внесенное мнение или оценка не должны

пропадать, ни «случайно», ни по сознательному умыслу несогласных пользователей из состава администрации системы. Система должна стимулировать открытый и честный обмен информацией, в том числе оценочными суждениями – для этого предусмотрен механизм «ссылок на критику». Выше мы уже обозначили информационную политику системы формулой «присутствуют и доступны все точки зрения».

3. Ясно, что механизм критических ссылок не может обеспечить защиту от внесения заведомо недобросовестной информации, поскольку он рассчитан как раз на добросовестных авторов. Мы полагаем, что обеспечить отбор добросовестных авторов можно, например, таким же способом, который применяется в электронном архиве научных препринтов Стенфордского университета [5]. Суть этого способа в следующем. Исторически, архив создавался с целями, похожими на предлагаемую систему общественных банков данных: как система максимально облегченного доступа студентов университетов к передовым научным публикациям до их официальной публикации в рецензируемых профильных журналах. Сначала режим размещения препринтов был полностью свободным, но со временем администрация архива ввела механизм *индорсмента* (endorsement). Индорсмент означает, что для первичного размещения статьи в архиве (новым автором) нужно получить разрешение «проверенного» автора архива. В этом случае аудитория пользователей системы разбивается (как уже предполагалось выше) на две категории: «читатели» и «писатели», причем круг «писателей» постоянно расширяется путем кооптации в него новых авторов самими «писателями». Можно подумать о том, чтобы сделать этот механизм более гибким и демократичным за счет разрешения на публикацию для всех, но с большим периодом между созданием записей. Этот период может сокращаться по мере получения «молодым» автором положительных отзывов, как от «читателей», так и от «писателей». В этом случае, возможно, потребуется

расширить описанную выше базовую структуру единичной записи числовым полем «оценка материала», которое могут изменять (на +\- 1) не только «писатели» - критики, но и «читатели».

4. С точки зрения вопроса хранения данных можно рассмотреть различные сценарии физического хранения данных. Это может быть централизованный сервер, облачные системы или распределенное хранение типа торрент-систем [6].

Литература

1. G3-Россия. Сетецентрическая система управления страной: http://www.viphmn.ru/index.php?option=com_content&view=article&id=233:-qg3-q&catid=7:2010-12-16-11-40-09&Itemid=37.

2. *В. И. Антипов, И. В. Десятов, Г. Г. Малинецкий, П. Л. Отоцкий, В. В. Шишов*. «Центр внедрения технологий социально – экономического планирования в России и прогнозирования мировой динамики». Препринт ИПМ им. М. В. Келдыша РАН №10 за 2009 г.

3. Хештеги – статья в Википедия: https://ru.wikipedia.org/wiki/%D0%A5%D0%B5%D1%88%D1%82%D0%B5%D0%B3

4. Википедия: Как создать статью/Свод правил. https://ru.wikipedia.org/wiki/Википедия:Как_создать_статью/Свод_правил.

5. Arxiv.org, статья в Википедия: https://ru.wikipedia.org/wiki/ArXiv.org.

6. BitTorrent, статья в Википедии: https://ru.wikipedia.org/wiki/BitTorrent.