Индексирование фактографических данных в документально-фактографических базах

ЛИХАЧЕВ Ю.М.

Институт черной металлургии НАН Украины

Представлен метод построения тезауруса для фактографических данных документально-фактографической базы физико-химических свойств шлаковых расплавов.

The method of construction of thesaurus is presented for factual data information of document - factual data base of physical and chemical properties of slag fusions.

Наведений метод побудови тезаурусу для фактографічних даних документально-фактографічної бази бази фізико-хімічних властивостей шлакових розплавів.

Введение. Проблема поиска и использования информации [1] – одна из самых актуальных в современной науке. Использование информации базируется на преобразовании ее в такую форму, которая позволяет более удобное и оперативное ее извлечение для эффективного использования в научной и производственной деятельности.

Выбор источников информации, стратегий ее поиска, способов оценки достоверности и соответствия уже имеющимся сведениям, методы оценки полезности найденной информации для разрешения соответствующих проблем – все эти факторы являются характеристиками информационного поведения человека.

Критериями информационной культуры человека можно считать его умение адекватно формулировать свою потребность в совокупности информационных ресурсов, перерабатывать, создавать качественно новую, адекватно отбирать и оценивать информацию.

Постановка задачи. Информационная система (ИС) служит для сбора и накопления информации, её эффективного использования для различных целей.

Информационные системы подразделяются на фактографические, документальные и документально-фактографические.

Фактографические ИС накапливают и хранят данные в виде множества экземпляров одного или нескольких типов структурных элементов (информационных объектов). Каждый из таких экземпляров структурных элементов или некоторая их совокупность отражают сведения по какому-либо факту, событию и т.д., отделенному от всех прочих сведений и фактов (отсюда и название – фактографические).

В документальных ИС единичным элементом информации является не расчлененный на более мелкие элементы документ и информация при вводе (входной документ), как правило, не структурируется, или структурируется в ограниченном виде.

Документально-фактографические (Смешанные) ИПС – это системы, которые занимают промежуточное положение между документальными и фактографическими ИС. Хранимая в них информация имеет вид документов, но по своей природе эти документы чаще всего частично структурированы, что существенно облегчает извлечение из них информации.

Информационно-поисковые системы ориентированы на извлечение подмножества хранимых сведений, удовлетворяющих некоторому поисковому критерию. Причём пользователей интересует не столько результаты обработки этих сведений, сколько сама извлекаемая информация.

В автоматизированных ИПС используются Информационно Поисковые Языки дескрипторного типа. Дескрипторы – это термины естественного языка, выражающие определенные понятия. Словарь дескрипторов с указанными между ними смысловыми отношениями, охватывающий определенную область знания, называется информационно-поисковым тезаурусом (ИПТ) [2]. Идея применения ИПТ для информационного поиска документов состоит в описании содержания документов и запросов с помощью дескрипторов, входящих в его состав. В тексте документа, вводимого в ИПС, выделяются слова, несущие основную смысловую нагрузку, так называемые ключевые слова (КС). При помощи ключевых слов достаточно точно передается содержание документа. После этого каждое слово заменяется близким ему по смыслу дескриптором информационно-поискового тезауруса.

Тезаурус представляет собой словарь, отображающий семантические отношения между лексическими единицами дескрипторного информационно-поискового языка (дескрипторами) и предназначенный для поиска слов по их смысловому содержанию.

Построение тезауруса состоит из нескольких взаимосвязанных этапов.

Первый этап - формирование словника. Словник - первоначальное множество характерных слов рассматриваемой предметной области. При этом рассматривается представительный массив наиболее информативных для данной предметной области документов. Выбираются слова, употребляемые в источниках, при этом учитываются частота употребления слов.

Второй этап - формирование множества ключевых слов. Из словника формируется множество ключевых слов. При отборе ключевых слов учитывается информативность слова, которая определяется исходя из частоты встречаемости слова, роли слова в данной предметной области. Третий этап - формирование классов эквивалентности. Выделение дескрипторов.

Совокупность терминов тезауруса - дескрипторов, заменивших ключевые слова, образует поисковый образ документа (ПОД). Точно так же на язык дескрипторов переводится и запрос.

Разработанная в ИЧМ база данных “Шлак”[3] на основе ИПС “BaseReader”(ИЧМ), построена по документально-фактографическому принципу хранения документа и содержащая более 230 документов и 8000 химсоставов, позволяет находить описания эксперимента, химического состава и свойств шлаков. Для поиска в документальной ИС используется индексирование по-

© Лихачев Ю.М., 2010

лей (Система, Добавки, Авторы). Такой способ позволяет выделить все релевантные документы для конкретного запроса. Обычно документальная информация используется для просмотра и поиск по индексированным полям не представляет проблем.

Фактография документа (рис.1), содержит структурные части (фактографические таблицы) – шлак, вязкость, электропроводность, поверхностное натяжение. При запросе по индексированным полям фактографическая часть не изменяет релевантность. Однако, при запросе по компонентам (Ключевые слова : CaO, Al2O3, SiO2, N140, G150 …), объем выборки (количество документов) может быть настолько огромным, что для его последующей обработки в прикладных программах приходится приложить немало усилий.

Для уменьшения объема выборки и повышения релевантности запроса был создан поисковый тезаурус. Формирование словника выполнялось на основе выборки из базы по фактографическим данным (рис. 2) и индексирования семантики определяемой пользователем полей “Система” и “Добавки”, включающий реляционное дерево запросов по химсоставу и свойствам (рис. 3).

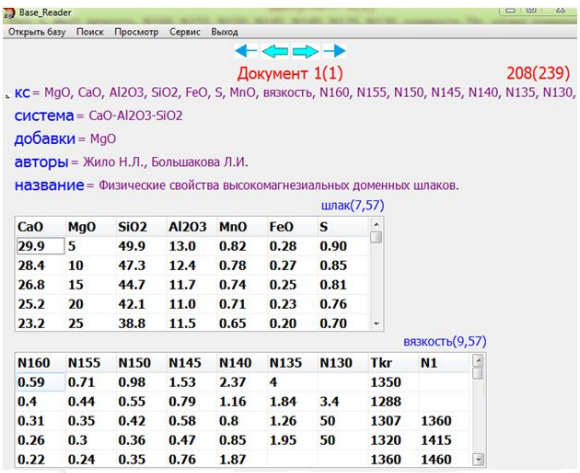

Рис. 1. Запрос к базе данных “Шлак”

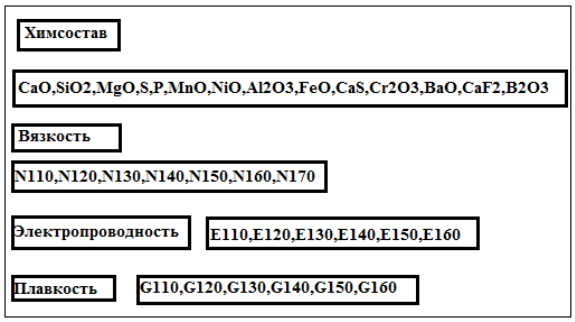

Рис. 2. Формирование словника

Запрос для фактографии может выполняться как совместно с документальным поиском, так и непосредственно по фактам. Для детализации фактографического поиска вначале выполняется документальный поиск по индексированным полям, на следующем этапе из найденных документов выбираются факты с учетом поискового фактографического тезауруса. Результат получается по нисходящей иерархии (Авторы, Система, фактография). Для поиска только по фактографии результат запроса выборка выполняется по всей базе.

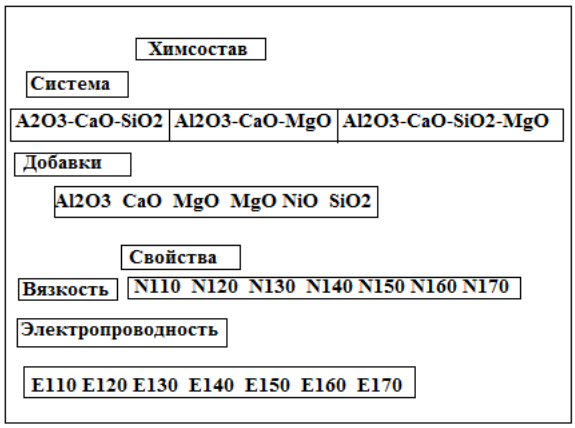

Рис. 3. Построение фактографического тезауруса.

Выводы

Таким образом, предложенный способ построения фактографического тезауруса и использования его для уменьшения объема выборки и повышения релевантности запроса позволяет оптимизировать процесс получения информации для комфортной работы по экспертной оценке данных [4] и разработке полуэмпирических моделей по прогнозированию физико-химических и технологических свойств металлургических шлаков – плавкости, поверхностного натяжения, энтальпии, электропроводности, серопоглотительной способности, минералогического состава [5].

ЛИТЕРАТУРА

1. Н.А. Гайдамакин. Автоматизированные информационные системы, базы и банки данных, М.: «Гелиос», 2002.
2. Черный А. И., Общая методика построения тезаурусов, "Научно-техническая информация. Сер. 2", 1968.
3. Тогобицкая Д.Н., Хамхотько А.Ф., Лихачев Ю.М. Оптимизация металлургических технологий и концепция создания информационно-интеллектуальных систем. //Фундаментальные и прикладные проблемы черной металлургии. Сб. науч.тр. -Киев. Наукова думка. –1995. – С. 242-249.
4. Лихачев Ю.М.б Тогобицкая Д.Н. , Хамхотько А.Ф. Экспертная система оценки достоверности экспериментальных данных о свойствах металлургических составов.//Математичне моделюваня.2(17) 2007.-С. 94-98.
5. Приходько Э.В., Тогобицкая Д.Н. Комплексное использование методологии физико-химического и математического моделирования для оптимизации металлургических технологий. //Сб.н.т. ИЧМ «Фундаментальные и прикладные проблемы черной металлургии».–Днепропетровск. – Вып.19.– 2009.– С. 368-377.

пост. 19.05.2010