Автоматический авторитетный контроль для распределенных библиографических баз данных

А.А.Князева
Отдел проблем информатизации ТНЦ СО РАН
amili@mail.ru

И.Ю. Турчановский
tur@hcei.tsc.ru

О.С. Колобов
okolobov@hcei.tsc.ru

Институт сильноточной электроники СО РАН

В работе представлен алгоритм автоматического авторитетного контроля, позволяющий связывать библиографические и авторитетные записи. Анализ факторов, влияющих на соответствие записей и расчет параметров алгоритма производились с помощью статистического эксперимента на библиографических и авторитетных базах данных. Описанный подход можно применять для решения задачи слияния дублетных записей (как библиографических, так и авторитетных).

1. Введение

В настоящее время большинство крупных библиотечных каталогов формируется с применением технологии авторитетного контроля записей [1]. Это позволяет упростить работу каталогизаторов и других специалистов, а также повысить качество записей. На практике, при объединении двух и более библиотечных каталогов, неизбежно приходится сталкиваться со следующими ситуациями:

1. необходимость выявления и слияния дублетных библиографических записей;
2. библиографические записи на материалы одного и того же автора, содержат разные контрольные номера авторитетной записи(или совсем их не содержат). Это следствие того, что для каждого библиотечного каталога применяется свой набор авторитетных файлов.

В отечественной практике нет примеров (точнее нам они неизвестны), которые бы показали как можно разрешать эти ситуации без участия человека (или точнее с минимальным участим человека). С другой стороны задача объединения библиотечных каталогов достаточно часто встречается на практике и влечет за собой необходимость технологии, которая позволила бы успешно разрешить указанные проблемные ситуации.

В данной работе исследовалась вторая проблема, т.е. анализировалась возможность создания технологии автоматического авторитетного контроля. Следует отметить, что описываемый подход может быть перенесен и на первую ситуацию, что позволит разработать технологию выявления дублетных записей и их слияния.

2. Обзор

За рубежом задача слияния дублетных авторитетных записей решается в рамках проекта VIAF (виртуальный авторитетный файл) Международной федерации библиотечных ассоциаций и учреждений (ИФЛА). Целью проекта VIAF является обеспечение возможности автоматического сопоставления и связывания авторитетных записей из различных национальных авторитетных файлов [2].

3. Теория

В основе предлагаемого подхода лежит принцип дискриминации пары «авторитетная запись – библиографическая запись» на два класса: соответствующих и несоответствующих пар. В качестве обучающей выборки для дискриминации выступают такие пары записей, про которые известно, к какому из двух классов они относятся. Определив эти два класса и рассчитав их центроиды (в качестве центроида использовалось среднее по классу), можно отнести любую новую пару к тому из классов, к которому она окажется «ближе» и тем самым принять решение о соответствии либо несоответствии записей друг другу.

Для того, чтобы оперировать понятиями ближе-дальше было выбрано расстояние Махаланобиса [3, 4], характеризующееся тем, что оно учитывает корреляции между переменными и является инвариантным к масштабу. Кроме дискриминации пар записей, с помощью расстояния Махалонобиса можно произвести отбор наиболее информативных признаков (то есть признаков, по которым классы разделяются наиболее четко).

Факторные переменные, используемые в работе, измеряются в интервальной шкале, что позволяет вычислять такие статистические характеристики как среднее и ковариация, однако не подчиняются нормальному распределению, что исключает применение параметрических критериев. Для проверки гипотезы значимости различия в средних по группам использовался критерий Хи-квадрат. Переменные, для которых принималась гипотеза об отсутствии различий (при уровне значимости 0,01), исключались из работы.

4. Алгоритм

В общих чертах алгоритм автоматического авторитетного контроля выглядит следующим образом:

1. В момент загрузки библиографической записи в базу данных анализируется поле, в котором указан автор источника и находятся все записи из авторитетной базы данных, с такими же ФИО;

2. Для каждой пары «авторитетная запись – библиографическая запись»:

   (a) рассчитываются значения факторных переменных и расстояния до классов (соответствующих и несоответствующих пар);

   (b) на основе того, к какому классу пара ближе принимается решение о ее соответствии либо несоответствии

3. В случае, если соответствующей признана ровно одна пара в библиографической записи указывается код авторитетной записи, если больше одной - запись отправляется на дополнительный контроль с привлечением специалистов, в противном случае никаких отметок не делается и запись попадает в базу.

Описанный алгоритм можно применять и для записей, которые уже хранятся в базе данных, но не получили контрольных номеров соответствующих авторитетных записей.

5. Эксперимент

Статистический эксперимент проводился на системе, включающей:

1. Библиографическую базу данных, около 300000 записей;

2. Базу данных авторитетных файлов на авторов, около 10000 записей.

На основе этих баз данных случайным образом были составлены обучающая (421 пара записей) и тестовая выборка (624 пары) из авторитетных записей однофамильцев с одинаковыми инициалами и библиографических записей, для которых были известны контрольные номера авторитетных записей. На основе обучающей выборки были рассчитаны статистические параметры алгоритма, после чего алгоритм был применен к тестовой выборке. Неправильный прогноз был дан всего для 2 пар из 624, причем при ближайшем рассмотрении в ошибочно классифицированных записях оказались неправильно указанные контрольные номера авторитетных файлов, то есть на самом деле и эти записи были классифицированы верно.

Таким образом, в результате эксперимента были получены достаточно обнадеживающие результаты, позволяющие утверждать, что рассматриваемый подход имеет право на существование. Для адекватной работы алгоритма необходимо периодическое уточнение его параметров с ростом базы данных. Кроме того, устойчивость алгоритма к ошибкам дискриминации можно повысить за счет разработки методики привлечения не использованной информации, например, информации о соавторах и тематике работы автора.

Список литературы

[1] Ковалева А.М. Авторитетный файл Имя лица /А.М. Ковалева // Библиотечное краеведение в информационном пространстве региона Барнаул, 2008. – С. 172–178.

[2] Bennett, Rick, Christal Hengel-Dittrich, Edward T. O'Neill, and Barbara Tillett. 2007. "VIAF (Virtual International Authority File): Linking the Deutsche Nationalbibliothek and Library of Congress Name Authority Files."International Cataloging and Bibliographic Control 36,1: 12–19.

[3] Факторный, дискриминантный и кластерный анализ: Пер. с англ./Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. – М.: Финансы и статистика, 1989. – 215 с.

[4] Афифи А., Эйзенс С. Статистический анализ. Подход с использованием ЭВМ. Пер. с англ. – М.: Мир, 1982. – 488 с.