Автоматический авторитетный контроль для распределенных библиографических баз данных

Князева А.А., Турчановский И.Ю., Колобов О.С.

Отдел проблем информатизации ТНЦ СО РАН

Институт сильноточной электроники СО РАН

Основные проблемы при слиянии каталогов

1. Дублетные записи – две и более записи (как авторитетные, так и библиографические) на один источник из разных библиотек

Основные проблемы при слиянии каталогов

1. Дублетные записи – две и более записи (как авторитетные, так и библиографические) на один источник из разных библиотек
2. Записи на материалы одного и того же автора, содержат разные контрольные номера авторитетной записи (или совсем не содержат их)

Обзор

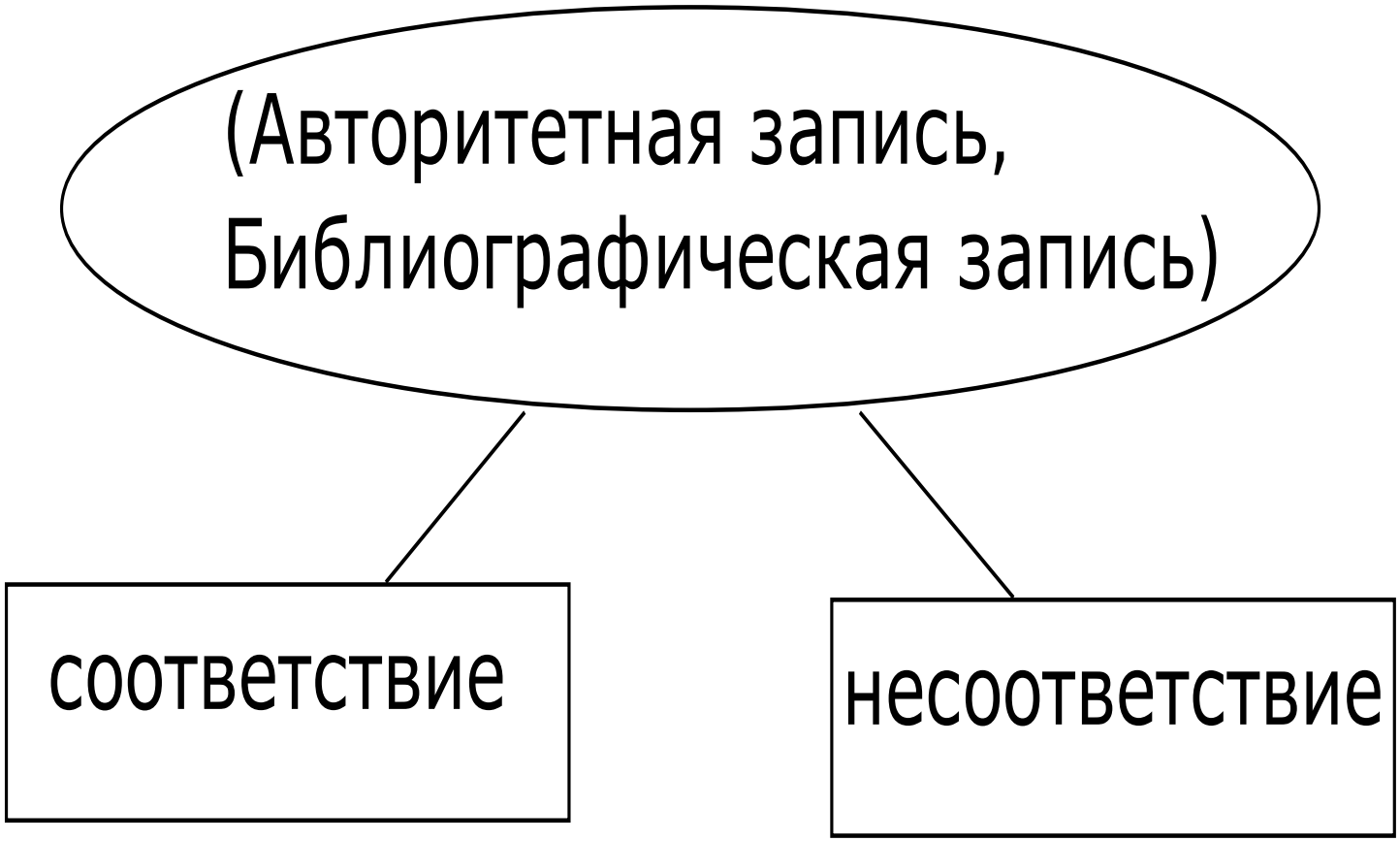
Проект **VIAF**(виртуальный авторитетный файл) Международной федерации библиотечных ассоциаций и учреждений (ИФЛА) - совместный проект OCLC, Библиотеки конгресса, Немецкой национальной библиотеки и национальной библиотеки Франции для разработки виртуальной комбинации имен авторов из каждого института в единый авторитетный сервис.

В настоящий момент в проекте участвует 15 организаций и обрабатывается 18 различных типов авторитетных записей.

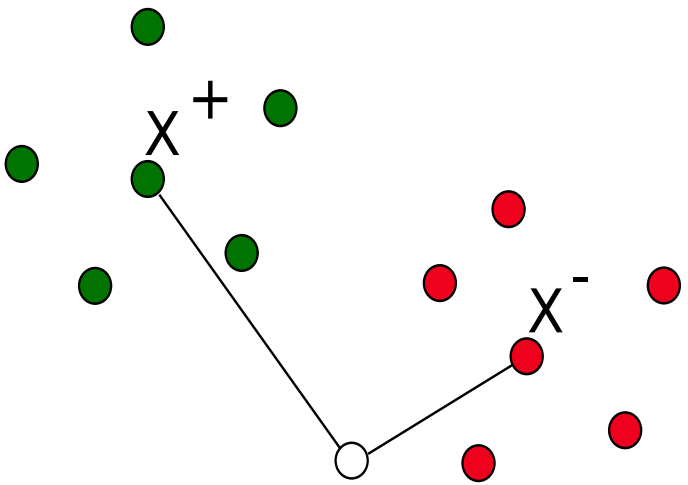
Постановка задачи

Создание алгоритма для связывания библиографической записи с соответствующей авторитетной записью (автоматического авторитетного контроля).

Задача связывания сводится к задаче дискриминации



Модель



1. Факторные переменные
2. Центроиды X^+, X^-
3. Расстояние
4. Обучающая выборка

Обучающая и тестовая выборки

Коллекции, предоставленные НП МедАрт:

1. Библиографическая БД, около 300000 записей
2. Авторитетная БД имен лиц, около 10000 записей

Всего было составлено 1045 пар записей

1. Обучающая выборка - 421 пара
2. Тестовая выборка - 624 пары

Факторные переменные

Переменная	Значение	Код	Комментарий	АЗ	БЗ
birth	не совпадает	1	совпадение с точностью до года	200$f	701$f
	не указана дата	2			
	совпадает	3			
death	не совпадает	1	совпадение с точностью до года	200$f	701$f
	не указана дата	2			
	совпадает	3			
addition	нет совпадений	1	совпадение усеченных форм	200$c	701$c
	не указано	2			
	одно совпадение	3			
	два совпадения	4			
	...	...			

Факторные переменные - продолжение

Перемен-ная	Значение	Код	Коммента-рий	АЗ	БЗ
place1	не совпадает	1	совпадение усеченных форм	200$y	712$c
	не указано место	2			
	совпадает	3			
place2	нет совпадений	1	вхождение усеченных форм	200$y	712$a
	не указано место	2			
	хотя бы одно совпадение	3			
work 1	не совпадает	1	совпадение усеченных форм	830$a	701$p
	не указано место	2			
	совпадает	3			
work2	нет совпадений	1	вхождение усеченных форм	830$a	712$a
	не указано место	2			
	хотя бы одно совпадение	3			

Фрагмент исходных данных

out	Факторные переменные						
	addition	birth	death	place1	place2	work1	work2
+	3	3	2	3	1	1	3
+	1	3	2	2	2	2	2
-	1	1	2	3	2	2	2
...							

1 - информация не совпала;
2 - нет информации;
3 - информация совпала.

Переменная	Значение	Код	Коммент.	АЗ	БЗ
out	не соответствует	-	точное совпадение	001	701$3
	соответствует	+			

Анализ средних значений

Переменная	X^-	X^+	p-value	Разница в средних
addition	1,122	1,924	$<2{,}2*10^{-16}$	значима
birth	1,076	2,912	$<2{,}2*10^{-16}$	значима
death	2	2,063	$1{,}6*10^{-4}$	мало значима
place1	1,855	2,151	$3{,}3*10^{-10}$	значима
place2	1,786	1,874	0,01216	незначима
work1	1,794	1,937	$1{,}2*10^{-4}$	мало значима
work2	1,786	1,899	0,0014	мало значима

Расстояние Махаланобиса

$$D^2(X, Y) = (X - Y)W^{-1}(X - Y)^T, \tag{1}$$

W - внутригрупповая матрица ковариации:

$$W_{ij} = \frac{1}{n. - 2} \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.}), \tag{2}$$

g - число классов;

n_k - число наблюдений в k-м классе;

$n.$ - общее число наблюдений по всем классам;

X_{ikm} - величина переменной i для m-го наблюдения в k-м классе;

$X_{ik.}$ - средняя величина переменной i в k-м классе.

Пример классификации

Факторные переменные						Расстояния		Прогноз класса	Значение out
addition	birth	death	place1	work1	work2	$D^2(X, X^-)$	$D^2(X, X^+)$		
3	3	2	3	1	3	124,5	27,8	+	+
1	3	2	2	2	2	65,6	2,6	+	+
1	1	2	3	2	2	7,8	76,9	-	-
...									

Проверка на тестовой выборке

Набор переменных	Классифицированы			
	ошибочно		правильно	
birth	13	2,1%	611	97,9%
birth, work2	2	0,3%	622	99,7%
birth, work2, addition	2	0,3%	622	99,7%
birth, work2, addition,...,work1	2	0,3%	622	99,7%

Алгоритм автоматического авторитетного контроля

1. В момент загрузки библиографической записи находятся все записи из авторитетной БД, с такими же ФИО;
2. Для каждой пары «авторитетная запись – библиографическая запись»:
   - 2.1 рассчитываются значения факторных переменных и расстояния до центроидов обоих классов ;
   - 2.2 принимается решение о соответствии или несоответствии записей.

Фрагмент библиографической записи

00161/H340-682478

700 1$a Шилов $b Б. В.$g Борис Владимирович$c цитолог
$f 19710323 $3 AShilov_BoriB200310066348070O

701 1 $aИванов$b В. В. $g Владимир Владимирович $c
биохимик
$f 19530130 $3 AIvanovVladV200404296348070O
$p кафедра биохимии и молекулярной биологии СГМУ

701 1 $a Казанский $b В. Е.

71202 $a Сибирский медицинский университет $c Томск
...

Фрагмент авторитетной записи

001 AIvanovVladV2004042963480700

200 1$a Иванов $b В. В. $c биохимия $f 19530130
$g Владимир Владимирович $y Томск

830 $a Образование: в 1975 г. окончил Томский университет, биолого-почвенный факультет, аспирантуру в Томском медицинском институте.
$a Ученая степень: в 1975 г. защитил кандидатскую диссертацию.
Кандидат биологических наук.
...

Требования к записям

Для авторитетных:

1. Наличие однофамильцев

Требования к записям

Для авторитетных:

1. Наличие однофамильцев
2. Требование полноты - наличие полей 200($a, $b, $c, $f, $y) и 830$a

Требования к записям

Для авторитетных:

1. Наличие однофамильцев
2. Требование полноты - наличие полей 200($a, $b, $c, $f, $y) и 830$a

Для библиографических:

1. Указание на авторитетную запись (наличие поля 701$3)

Требования к записям

Для авторитетных:

1. Наличие однофамильцев
2. Требование полноты - наличие полей 200($a, $b, $c, $f, $y) и 830$a

Для библиографических:

1. Указание на авторитетную запись (наличие поля 701$3)
2. Требование полноты - определены значения как минимум двух переменных

Особенности предлагаемого подхода

Для работы с другими коллекциями записей необходим этап обучения для статистического вычисления параметров алгоритма.

Дальнейшее развитие

1. Дополнение алгоритма за счет привлечения новой информации

Дальнейшее развитие

1. Дополнение алгоритма за счет привлечения новой информации
2. Применение описанного подхода к задаче выявления дублетных записей

Автоматический авторитетный контроль для распределенных библиографических баз данных

Князева А.А.

Отдел проблем информатизации ТНЦ СО РАН

amili@mail.ru

Спасибо за внимание!