Современные проблемы дистанционного зондирования Земли из космоса.
2012. Т.9. №3. С. 38–42

Использование метода нечеткого поиска для идентификации судов, по их атрибутам в разнородных БД

В.В. Марченков [1], В.Н. Пырков [1], В.Н. Черных [1], А.В. Солодилов [2], В.В. Ермаков [3]

[1] 1Институт космических исследований РАН (ИКИ РАН),
117997, Москва, Профсоюзная, 84/32,
e-mail: pyrkov@d902.iki.rssi.ru
[2] ООО Федеральное государственное учреждение «Центр системы мониторинга рыболовства и связи» (ФГУ «ЦСМС»),
107996, Москва, Рождественский бул., 12
[3] ООО «Камчатские системы связи и мониторинга» (ООО «КССМ»),
683031, Петропавловск-Камчатский, Давыдова, 7,
e-mail: wwe@mail.ru

В статье приведен пример использования технологий нечеткого поиска для решения задачи оценки числа судов рыболовного флота России, оборудованных автоматизированными идентификационными системами (АИС). Анализу были подвергнуты сведения, хранящиеся в разных базах данных (БД), содержащих частично совпадающую информацию о судах.

Ключевые слова: мониторинг, рыболовный флот, АИС, нечеткий поиск.

Введение

Традиционные системы глобального мониторинга судов используют данные, получаемые с искусственных спутников земли. Судно, оборудованное техническими средствами контроля, в которые входит приемник GPS/ГЛОНАСС, передает по расписанию информацию о своем местоположении по каналам спутниковой связи. Полученная информация накапливается в БД. Кроме позиционных данных в БД системы мониторинга накапливается информация о судне (год постройки, владелец, размеры судна, мощность двигателей, имеющееся оборудование и т. д.). При мониторинге рыбопромысловых судов в БД также поступают ежедневные отчеты судов о выполняемых ими действиях. К этой информации добавляются сведения о выданных разрешениях на добычу биологических ресурсов. По этому принципу организована БД отраслевой системы мониторинга судов рыбопромыслового флота Росрыболовства РФ (БД ОСМ).

В последнее десятилетие была разработана технология глобального мониторинга морских судов Satellite Automated Identification System (SatAIS) — наблюдение за сигналами автоматизированных идентификационных систем (АИС) с помощью спутника. Коммерческое использование этой системы стартовало в начале 2011 г.

Система разрабатывалась с целью повышения безопасности мореплавания в условия большой концентрации судов (в портах, при прохождении проливов и каналов), как средство предупреждения столкновений судов. В настоящее время многие суда обязаны иметь на борту это оборудование. В систему входят судовые и наземные приемо-передатчики, работающие в диапазоне ФМ. Судовые передатчики передают информацию о координатах судна, его скорости, курсе, габаритах и т. д. Береговые передатчики передают информацию о погоде, поправки для корректировки показаний GPS-приемника, указания судам о смене курса и т. д.

Работу системы обеспечивает программное обеспечение, которое организует из всех работающих передатчиков сеть, отслеживает изменение координат и курса судна, формирует сообщения, которыми обмениваются приемо-передатчики АИС, выбирает момент для передачи своего сообщения и приема входящих. Другими словами, приемо-передатчики АИС по

определенному алгоритму выбирают одну из двух частот и момент, в который они передают информацию. Таким образом, группа судов образует локальную сеть из передатчиков, находящихся в зоне радиовидимости друг друга (у поверхности земли она почти совпадает с границей прямой видимости), которая обеспечивает возможность обмена информацией между судами. Максимальное расстояние, на котором суда могут обмениваться информацией, ограничивается кривизной поверхности Земли и высотой размещения антенны передатчика АИС, что составляет около 20 миль.

Сигналы передатчиков АИС удается принимать из космоса, что позволяет использовать АИС и как систему глобального мониторинга. Спутник одновременно собирает информацию с территории диаметром 5000 км.

Очевидно, что новая система определения позиций судов SatAIS может использоваться наряду с традиционными методами позиционирования системами мониторинга судов, в связи с чем рассмотрим задачу слияния разных информационных потоков в для решения конкретных задач. Например, практическую ценность может представлять оценка доли судов рыболовного флота, оборудованных в настоящее время аппаратурой АИС. Подобные сведения полезны при проведении оценок экономической целесообразности использования технологий SatAIS в системе мониторинга.

Постановка задачи

Была поставлена задача оценить количество судов из числа зарегистрированных в БД ОСМ, имеющих на борту оборудование АИС. Сведения о судах, оснащенных АИС, были получены с сайта http://aprs.fi, накапливающего данные с помощью расположенных на берегу приемников АИС. Сравнение атрибутов этой информации и сведений о судах, хранящихся в БД ОСМ, показало, что они пересекаются по полям: радиопозывные судна и наименование судна. К сожалению, практика такова, что радиопозывные не являются однозначным идентификатором судна. При внимательном рассмотрении выяснилось, что и названия судов тоже не очень надежный параметр — суда могут носить совпадающие названия. Существуют и другие источники искажения информации. Например, в рассматриваемом случае названия судов вносятся в память оборудования АИС только латинскими буквами, а это плохо стандартизировано. Мало того, что можно несколькими способами транслитерировать русские буквы в латинские, не редко встречаются сокращения одного из слов, входящих в название, дополнение название (к названию добавляют порт приписки), возможна замена римских цифр на арабские и наоборот, использование цифр вместо сходных по написанию букв и т. д. Однако одновременное совпадение названия судна и радиопозывных должно быть достаточно надежным признаком того, что речь идет об одном и том же судне. Во всяком случае, нами было выявлено несовпадение радиопозывных и названия судна приблизительно в 0,8 % случаев из числа зарегистрированных в БД ОСМ, и ни одного случая совпадений имени и радиопозывных для разных судов.

Если принять, что судно идентифицировано, когда у него совпали и радиопозывные, и название, то исходная задача выявления судов рыболовного флота, оборудованных АИС, сводится к проблеме сравнения двух вариантов названия судна в некотором «нечетком» смысле. На сайте API [http://aprs.fi] мы отобрали сведения о 386 судах, зарегистрированных под флагом России, и заявившими, что занимаются рыболовным промыслом. Из БД ОСМ были отобраны названия судов, чьи радиопозывные совпали с радиопозывными отобранных для анализа судов. Был сформирован массив данных, содержащих два варианта названий судов латинскими буквами (по версии, предоставленной сайтом, и по версии БД ОСМ) и набор остальных идентификаторов из обоих источников.

Методика

Для сравнения степени похожести двух строк мы использовали алгоритмы, обычно используемые в задачах, связанных с нечетким поиском. Нечеткий поиск подразумевает введение некоторой метрики, которая принимается за меру подобия строк. Для вычисления меры совпадения строк мы применили алгоритм, носящий имя Левенштайна, впервые упомянувшего задачу нечеткого сравнения в 1965 г. (Левенштайн, 1965). Расстояние Левенштайна — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую. Поскольку длина строк, составляющая названия судов, разная, то применялась нормировка полученной длины по следующей формуле:

$S = 1 – L/LSS,$

где S – коэффициент похожести строк, далее просто отношение; L — расстояние Левенштайна между сравниваемыми строками; LSS — суммарная длина сравниваемых строк.

При этом если $LSS = 0$, то $L = 1$.

При $S = 1$ наблюдается полное совпадение строк, $S = 0$ означает полное несовпадение.

В применении к нашей задаче методика выглядела следующим образом:

- выявление совпадений радиопозывных хранимых в БД ОСМ и в наборе данных полученных с передатчиков АИС;
- транслитерация для отобранных судов русских названий латинскими буквами;
- нахождения соответствий методом нечеткого поиска между парами названий, полученными через АИС и из БД, при различных значениях коэффициента похожести названий S;
- ручной анализ отобранных данных и выявление ошибок, т. е. неверной идентификации алгоритмом совпадения названий.

Алгоритм был реализован с помощью скрипта на языке Python. Кроме библиотек, входящих в стандартную поставку использовались библиотеки Levenshtein [http://pypi.python.org/pypi/python-Levenshtein/] — для проведения сравнения слов и MySQLdb [http://sourceforge.net/projects/mysql-python/] — для работы с базой данных.

В табл. 1 представлены данные о количестве ошибок, обнаруженных ручным анализом, и отобранном программой по интервалам для различных величин коэффициента похожести S.

Таблица 1. Сравнение результатов анализа совпадений строк полученных компьютером и человеком

Интервал значений коэффициента похожести S	**Число ошибок, выявленных ручным способом**	**Частота наблюдения значения S**
1,0–1,0	0	247
1,0–0,9	0	41
0,9–0,8	0	39
0,8–0,7	0	18
0,7–0,6	0	14
0,6–0,5	0	7
0,5–0,4	1	2
0,4–0,3	2	3
0,3–0,2	7	7
0,2–0,1	6	6
0,1–0,0	2	2

Можно отметить, что в области значений $S > 0,5$ ошибок нет. Далее при $0,5 > S \geq 0,3$ наблюдается фиксация ошибок и одновременно наблюдается минимум частоты наблюдения

строк в этих интервала коэффициента достоверности S. При более низких значениях S строки не совпадают.

Дискуссия

Результаты, представленные в табл. 1, позволяют сделать вывод, что при сравнении двух названий судов можно принять, что названия совпали, если значение $S > 0,5$. Наличие минимума числа пар строк с коэффициентом подобия в интервале $0,5 > S > = 0,3$ демонстрирует наличие достаточно четкой границы между совпадающими и несовпадающими названиями судов в терминах выбранного нами коэффициента похожести, что указывает на достаточную надежность этого критерия. Кроме того, если $S < 0,3$, то можно однозначно считать, что сравниваемые строки не совпадают.

Для того, что бы получить более четкое представления о выявленных различиях, рассмотрим более подробно сравниваемые строки в области $0,6 > S > 0,3$.

В табл. 2 представлены сами сравниваемые строки и значения S в интервале $0,6 > S \geq 0,5$, а в табл. 3 — в интервале $0,5 > S \geq 0,3$, т. е. когда ошибки еще не фиксируются и момент фиксации первых ошибок.

Таблица 2. Результаты сравнения в интервале $0,6 > S \geq 0,5$ (до фиксации ошибок)

Название судна по версии ОСМ (транслитерация русского названия)	Название судна по версии АИС	Значение коэффициента похожести
ALEKSANDR SHALIN	A.SHALIN	0,538
BOREY (MURM)	BOREY	0,588
TSEZAR	CAESAR	0,500
KATRIN (MURM)	KATHERINE	0,545
LIRA (KLG)	LIRA	0,571
NEREY (MURM)	NEREY	0,588
VEGA (MURM)	VEGA	0,533

Таблица 3. Результаты сравнения в интервале $0,5 > S \geq 0,3$ (впервые зафиксированы ошибки)

Название судна по версии ОСМ (транслитерация русского названия)	Название судна по версии АИС	Значение коэффициента похожести
VOLK ARKTIKI	ARCTIC WOLF	0,348
SEVERNYY OKEAN	NORTH OCEAN	0,480
PARNAS	PROMYSLOVIK	0,353
NIARA	STELLA KARINA	0,333
SOLOMON	SVIOLA	0,462

Как видно из данных, представленных в табл. 2 и 3, коэффициент похожести S действительно позволяет выявить границу, отделяющую названия, совпадающие точно, и названия, совпадающие условно, от заведомо несовпадающих. В табл. 2 попали «правильные» названия судов, но «дополненные» вторым словом, которое представляет сокращенное наименование порта приписки или название из двух слов с сокращением одного из слов до буквы, а также слово «ЦЕЗАРЬ» записанное в разных транскрипциях. Попавшие в табл. 3 данные нужно признать несовпадающими. Наряду с прямыми несовпадениями типа «NIARA» – «STELLA KARINA» и «SOLOMON» – «SVIOLA», мы наблюдаем случаи использования разных языков и транскрипций: «VOLK ARKTIKI» – «ARCTIC WOLF». Обнаруживаются ли эти названия совпадающими, решить не просто, потребуется консультация эксперта в данной предметной

области. Можно сказать, что при $S < 0,5$ мы наблюдаем значимые расхождения в названиях судов. Граница $S = 0,5$ достаточно надежно отделяет несовпадения названий судов от различий в их написании латинскими буквами и небольших вариаций названий в виде полного или сокращенного имени, добавления или опускания порта приписки судна и т. д.

Если принять, что ручной проверке подлежит только интервал $0,5 > S \geq 0,3$, то это позволит сократить долю ручного анализа со 139 случаев (разница между общим числом представленных для сравнения пар строк и числом полностью совпавших названий) до 5 (число пар, имеющих коэффициент похожести $0,5 > S > 0,3$), т. е. более чем в 27 раз. При этом компьютерный анализ позволяет надежно выявить 358 совпадений. Рассмотрение спорных случаев человеком увеличивает эту цифру до 360.

Выводы

Применение методов нечеткого поиска при сравнении названий судов хранящихся в разных базах данных:

- показало высокую эффективность;
- позволяет автоматизировать процесс объединения данные из разных информационных источников;
- облегчает выявление ошибочных данных при сравнении информации разных информационных источников;
- не менее 360 судов рыболовного флота имеют на борту оборудование АИС.

Литература

1. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. Академии Наук СССР. 1965. Т. 163. № 4. С. 845–848.

Application of the fuzzy search method to vessels identification using their attributes in heterogenous databases

V.V. Marchenkov [1], V.N. Pyrkov [1], V.N. Chernykh [1], A.V. Solodilov [2], V.V. Ermakov [3]

[1] Space Research Institute
117997, Moscow, Profsouznaya, 84/32,
e-mail: pyrkov@d902.iki.rssi.ru
[3] Federal State Department "The Centre of fishery monitoring and communication" (FSD "CFMC"),
107996, Moscow, Rogdestvenskiy bul., 12
[2] Kamchatka systems of communication and monitoring Ltd,
683031, Petropavlovsk-Kamchatka, Davidov, 7,
e-mail: wwe@mail.ru

This article shows how to use fuzzy search technology to solve the problem of estimating the number of equipped with AIS vessels of the RF fishing fleet. The information stored in different databases containing partially matching information about the vessels was analyzed.

Keywords: AIS, monitoring, fishing fleet, fuzzy search.