Анализ баз данных:

выявление ошибочных данных и восстановление недостающих

Любое предприятие располагает многочисленными данными, связанными с деятельностью предприятия. Иногда, эти данные довольно трудно регистрируемы, но в тоже время, эта операция является необходимой, так как они могут представлять огромную ценность для компании:

- На их основе можно, например, спрогнозировать объем продаж как по виду продукции, так и по сектору; расчитать необходимый или оптимальный объем производства и численность кадров; улучшить логистику.

- С их помощью представляется возможным выявлять проблемные ситуации, связанные с риском: в том или ином случае, если риск велик, всегда можно изменить тактику.

- Они предоставляют информацию акционерам, клиентам и общественности на фактической и неоспоримой основе.

Однако, несмотря на их стратегическое значение, базы данных очень часто несовершенны: об этом свидетельствует наличие множества ошибочных и недостающих данных. В худшем случае, такая совокупность данных теряет всякую надежность.

На протяжении многих лет SCM работает над развитием вероятностных методов, позволяющих улучшить информационную систему, а именно:

- выявление ошибочных данных;
- восстановление недостающих данных.

A. Выявление ошибочных данных

Процент ошибочных данных часто высок (более 10 %), и источники ошибок могут быть различными:

- Человеческий фактор: ошибки копирования во время регистрации данных (ошибки в единице, дате, поле), которые очень трудно контролировать.
- Измерительные приборы (плохая калибровка, недостаточная точность).

В ходе работы, которую мы осуществили в 2010 г. для французского агентства по атомной энергии (входящее в Организацию экономического сотрудничества и развития), мы выявили нарушения, которые могут касаться как одиночных ошибочных данных (их одной определенной особенности), так и какой-либо совокупности последовательных данных, тенденция которой явно отличается от тенденции всей совокупности данных. Мы разработали методы автоматического обнаружения ошибок, обеспечивающие предельно низкий процент погрешности.

Как только база данных будет проверена этими методами, пользователи смогут пользоваться этой информацией: они смогут видеть "показатели надежности", информирующие о качестве данных, и смогут выбирать, работать ли им со всей совокупностью данных или использовать только самые надежные данные.

B. Восстановление недостающих данных

Почти все базы данных имеют "пробелы", и причины этому могут быть разными:

- отсутствие человека, ответственного за измерение;
- поломка измерительного прибора;
- нехватка бюджета;
- удаление, потеря, износ и т.д.

В 2007 г. *Société de Calcul Mathématique* в рамках контракта с Veolia Environnement, провела работу по реконструкции базы данных, содержащей показатели ежедневного уровня воды в 19-ти реках департамента Вандея на западе Франции. Наблюдение за реками велось в течение 37 лет, а на момент начала нашего проекта недоставало 50% данных. Вероятностные методы, которые мы использовали для этой цели, подробно изложены в нашей книге "Вероятностные методы для восстановления недостающих данных".

Тем не менее, недостающие данные имеют положительный аспект: они позволяют экономить! Соответствующие методы, определяющие заранее, сбор каких данных может не проводиться и как совокупность данных может быть затем восстановлена, позволяют тем самым экономить на датчиках, измерениях и человеческих ресурсах.

Книга: Bernard Beauzamy et Olga Zeydina: Méthodes probabilistes pour la reconstruction de données manquantes. Ouvrage édité et commercialisé par la *Société de Calcul Mathématique SA*, ISBN : 2-9521458-2-2, ISSN : 1767 – 1175, avril 2007.

Во всех нижеперечисленных контрактах, выявление ошибочных данных и восстановление недостающих данных сыграли существенную роль:

- Veolia Environnement, 2005 : Analyse des pénuries d'eau en Vendée ;
- Agence Européenne de l'Environnement, 2006-2011 : Méthodes probabilistes pour la qualité de l'eau ;
- Veolia Environnement, Région Ouest, 2007 : Détection de dysfonctionnements dans les réseaux de capteurs ;
- Veolia Environnement, Région Ouest, 2007-2009 : Constitution d'un panel de consommateurs et prévision des consommations d'eau potable ;
- Institut de Radioprotection et de Sûreté Nucléaire, 2007-2011 : Applications de l'Hypersurface Pro-babiliste aux problèmes de sûreté des réacteurs nucléaires ;
- International Stainless Steel Forum, 2008 : Analyse générale du système d'information et préconi-sations relatives au traitement statistique des données ;

- Réseau Ferré de France, 2008 : Etude statistique concernant les causes des retards des trains en Ile de France ;
- Agence de l'Eau Artois-Picardie, 2008 : Etude probabiliste concernant la qualité des eaux de rivière et caractérisation des situations de bonne qualité ;
- Groupe Novalis, 2008 : Analyse critique de l'efficacité de certains dispositifs d'aide ;
- Snecma Propulsion Solide, 2009 : Méthodes probabilistes pour la fiabilité ;
- Caisse Centrale de Réassurance, 2009 : Etudes probabilistes relatives aux débits des rivières ;
- Fédération des Établissements Hospitaliers et d'Aide à la Personne, 2009 : Développement d'un système d'information ;
- Areva, 2010 : Méthodes probabilistes pour l'étude d'un stockage de déchets radioactifs ;
- Brigade des Sapeurs Pompiers de Paris, 2010 : étude statistique relative aux interventions ;
- Agence Nationale de l'Habitat, 2010 : Lois de probabilité relatives aux délais de paiement ;
- Nuclear Energy Agency (OCDE), 2010 : détection de données aberrantes dans les bases de données.