Center Parcs Deduplicating Internationally

David Hoogvorst

Center Parcs has six country-specific databases containing data on a total of 10 million customers. The data is frequently modified, and this means that database pollution must always be guarded against. An important aspect in the battle against pollution is the prevention of duplicates, and it is vital that deduplication is performed carefully so that customers are not offended in any way.

Introduction

The editor of this magazine asked me to write an article on handling large amounts of data in the case of customer relation systems.

Currently, I'm working as a technical product manager Intelligent Retrieval® for Human Inference, a company that settled in the niche market of enhancing the quality of relationship data.

Human Inference has three major product lines:

1. Name/IT®, a toolbox for validation, formatting and correction of names

2. itACA®, a toolbox for formatting and correction of addresses

3. Intelligent Retrieval®, a fuzzy identification method, that can be used for searching and deduplication purposes

In this article, I show how our products can help to prevent customers from 'drowning in data'.

Center Parcs

carefully deduplicating international data

Who hasn't heard of Center Parcs? Well, even if you've never visited one of their villages, you probably know somebody who has, or at least, somebody who has seen one of their brochures. Center Parcs is an internationally oriented company, and it maintains a huge international customer database. The customer data is constantly being updated and modified, and the prevention of database pollution is essential.

History

Piet Derksen, a Dutchman, started Sporthuis Centrum as a family business thirty years ago. The concept of a secluded holiday village with on-site sporting facilities was a unique concept at the time. Its great success soon generated a lot of competition. In the meantime, the name has changed to Center Parcs, but the concept is alive and well and the company is still the market leader. The following aspects characterise the Center Parcs concept:

- 'All-weather holidays'
- All facilities on-site in the holiday park
- Everybody completely free to do (or not to do) as they wish
- Always more on offer than you have at home (this used to mean cable TV in every cottage, now it means whirlpools and saunas in the Executive Villas!).

Center Parcs has villages in the Netherlands, Great Britain, Belgium, Germany and France (Table 1),

and since 1989 it has been in the hands of the British company Scottish and Newcastle. In terms of organisational structure, there are three autonomous divisions: Great Britain, France and a combination of Belgium, the Netherlands and Germany (BeNeDu for short).

Very carefully

The Center Parcs EDP department was started around 20 years ago, using a booking system the company designed itself. The system has since been rebuilt a number of times and it is optimised roughly every six months. Throughout the years, the customer database has grown exponentially, and regular cleansing is required to prevent pollution. During these cleansing sessions, data is also deleted. This must be done very carefully, however, because the company does not want to endanger the good relationship with its customers. After all, the mistaken deletion of booking data could lead to very dissatisfied customers indeed.

Ms Corma Otte (Client Database Corporate Manager) and Ms Carola Diederik (Project Coordinator/Information Analyst) are happy to explain the information architecture. There are, in fact, two systems:

1. LIMA, the *customer system*, in which bookings made within the last ten years are stored.

2. RES, the *reservation or booking system*, the operational component in which bookings made within the last three years are stored.

Using LIMA, Center Parcs manages more than 10 million customers internationally (names and addresses). The customer data is stored in six individual customer databases:

- Dutch
- British
- German
- Belgian
- French
- Other countries

Despite having a different databases for each geographical region, *all the data is internationally available.* This structure means, for example, that the data on an English customer in Germany who books a cottage in a Dutch holiday park will be stored in the UK database.

In LIMA, the following detailed data can be traced for any of the 5.4 million bookings:

- name and address data
- cost of the stay
- which holiday park was reserved
- composition of the group (children, adults, pets)
- applications for brochures (by telephone, as a result of surveys, coupons in magazines, etc.).

Deduplication process

"Modifications are often made in the reservation system as a result of bookings by telephone. Booking personnel perform an on-line check to see whether the customer already occurs in the database, and then check that data," relates Corma Otte. For this task, Intelligent Retrieval® ensures that customers who most closely match the one being checked are displayed on the screen. Although this way of working greatly reduces the occurrence of duplicates, batch deduplication is still used to ensure the database is kept as free from duplicates as possible. This is performed every night on all six of the country-specific databases, for both newly added and modified customer data. Corma Otte: "When we distribute the new brochures, we sometimes get peaks of 10,000 changes in one day." When there are too many records changed to process in one nightly run, the run is divided over several sessions. Intelligent Retrieval® determines the degree of similarity between any two records of customer data each time the process is performed. Center Parcs uses two threshold values that can differ for each country. If the results are above the upper threshold (and hence have a very high degree of similarity), the system deduplicates automatically. Cases that fall between the two thresholds are then dealt with manually. If necessary, staff will even look for information in the telephone directory, but it makes an address expensive. The "big deduplication" of all the customer data takes place continuously in cycles of 7 weeks, when the rewards for

Center Parcs — International facts and figures for 1998	
13 villages:	5 in the Netherlands
	3 in England
	2 in Belgium
	2 in France
	1 in Germany
Number of villas: (including hotels and apartments)	8,932
Turnover:	Around £300M
Number of guests:	over 3 million, consisting of:
	965,000 Dutch
	805,000 British
	577,000 Germans
	507,000 French
	260,000 Belgians
	5,000 from other nationalities
Number of overnight stays:	13.6 million
Percentage full in 1997:	85 to 90 %
Number of personnel:	9,400 (full-time and part-time)
Average length of stay:	77 % weekend or midweek
	21 % week
	2 % week+

Table 1: Center Parcs — International facts and figures for 1998

continuous checking are reaped - the last "big deduplication" generated just an average of 50 duplicates per 1 million customers.

Z-run

The so-called Z-run is unique to Center Parcs. Corma Otto: "Sometimes people live at one and the same address but the names are very different. In such cases, we don't want to deduplicate. Instead, the guest we haven't heard from the longest is classified as 'Z' and we treat the more recent guest as *the resident*. After deduplication, we do the Z-run. In this way, we avoid sending more than one brochure to an address, because the Z's are not included in the selections."

Deduplication European-style

The six separate country-specific databases each use their own version of Intelligent Retrieval® —

for example for the French database they run the French version of Intelligent Retrieval® and for the German database the German version of Intelligent Retrieval®. This means that the deduplication process is not the same in each country-specific database. Carola Diederik: "For example with the French database, we use two types of deduplication, making a distinction between rural and urban areas. This is necessary because the postcode density in cities is much greater and because addresses in France are not standardised. In the 'Other nationalities' database, every nationality is assigned a country code. Addresses with the same country code are deduplicated using a mixture of evaluations based on the French, German, English and Dutch versions of Intelligent Retrieval®."

Address authenticity percentage

Another measure used to safeguard the quality of personal data is the allocation of an address authenticity percentage. When a booking is made, an ad-

dress is given the maximum qualification of 70% correctness. Every six weeks, this percentage drops by one point. "People might have forgotten to tell us they've moved house," explains Corma Otte. Data obtained indirectly (for example, from list brokers) is given a score of 40% on the authenticity scale because it has not actually been verified by Center Parcs.

Traceability

The main function of the customer data system (LIM and RES) is the storage of customer data, but Center Parcs also uses it to analyse customer profiles, which are used to:

- perform direct mailing activities
- determine potential customers of specific package holidays (for example, tennis packages)
- compile customer-specific booking dialogues in the call centres.

Center Parcs is considering renovating or rebuilding the customer data system to make profile recognition simpler. The personal data of customers is becoming increasingly important and is now used in more and more areas. The ultimate aim is one-to-one marketing, for which the data must be in perfect condition. This is why Center Parcs is very careful with its customer data. Corma Otte: "We purchase very little data and certainly do not lease our databases to anyone else. When people ask us for brochures or other information, we always tell them what we'll be using their personal data for, and we always give them the option of not being included in our mailings in the future. Furthermore, we process returned mail as quickly as possible, and people who book through an agent are not included in mailings without the agent's express permission."

Carola Diederik: "Sometimes we get telephone calls from people who have received unsolicited mail. This is a particularly sensitive subject in Germany, for example. Occasionally we had lawyers asking us to reveal the sources of particular addresses. We can use Intelligent Retrieval® to trace data on people who are very similar to the person in question, and then explain the cause of the error. That also helps us to come across professionally to our customers."

Satisfied

Carola Diederik and Corma Otte are both "absolutely satisfied with the way Intelligent Retrieval® works." They add: "It's really flexible, and you can see this especially when making modifications to the software." The great advantage of Intelligent Retrieval® is that you can be as precise as you want with deduplication.

David Hoogvorst
david.hoogvorst@humaninference.com
www.humaninference.com