3.2.6 Matching

Peter van der Putten[1]

Matching supply and demand is a core process for industry and government, whether it concerns searching for products, services, persons or information. In this chapter we will give a quick introduction to the data mining technologies used for matching. Furthermore, we will illustrate this with two cases from the profit and non-profit sector, hunting for criminals and hunting for jobs. We will end with a vision of the future.

How does matching work?

Finding the best products to buy, selecting the best prospects to sell to, allocating the best matching resources — human or other — and serving the right information to citizens can all be seen as examples of matching processes. Given a demand, the best offers have to be found and evaluated (or the other way around). Typically, supply and demand do not correspond perfectly.

Classical information technologies, such as standard database queries, only offer a limited solution to this problem. If a user is not capable of formulating an exact, errorless query, a conventional query system is likely to return nonsense instead of a reasonable answer. Difficulties also arise when an exact query can be formulated, but no data satisfies all the specified criteria. The cause of these problems is that conventional search techniques rely on Boolean, all-or-nothing logic, which makes the search rigid and brittle: a fraction of an inch can make a world of difference. People are obviously able to handle fuzzy descriptions and match them to 'objects' in the real world. Conventional information technologies are hardly capable of doing this.

Nearest neighbor search

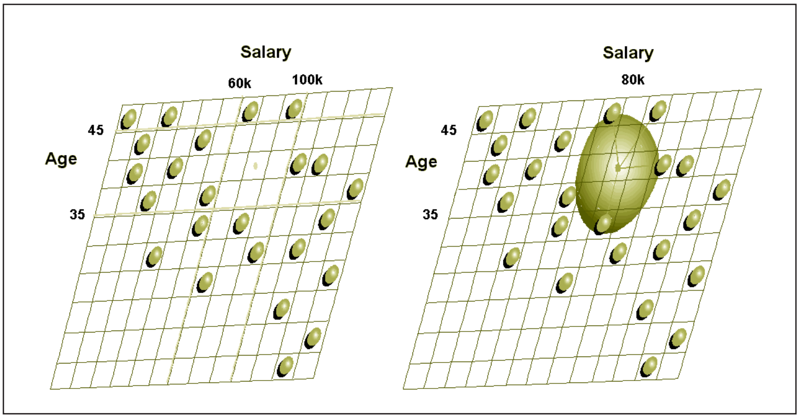
To support matching processes fully, alternatives should be found that match to a certain degree. Data mining offers a wide variety of technical solutions to support the matching process, including prediction and clustering. But the most basic and simple approach to this problem is nearest neighbor search (fuzzy matching).

The idea behind nearest neighbor matching can be explained with the following example from human resources matching (Figure 1). Assume we are looking for four persons that are around 40 years old with a salary of around 80,000 Euro. Each dot is a person in the database. If standard database queries are used (left), various strict search criteria should be set, for example: 35-45 years old and salary between 60,000 and 100,000 Euro. The user doesn't know before-

[1] Drs P. van der Putten, pvdputten@hotmail.com, pvdputten@liacs.nl, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. This section was written, while he was a consultant for Sentient Machine Research.

Figure 1

Matching as a data mining task: nearest neighbor search



hand how many persons fit the criteria perfectly (none in this case) and must consequently tune the constraints over and over again to get a selection of reasonable size and quality.

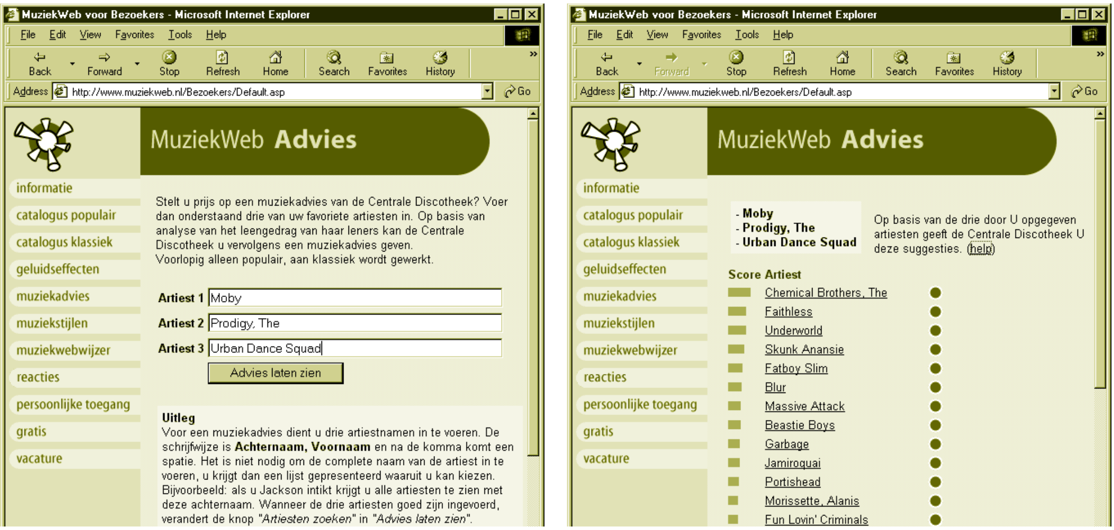
Nearest neighbor search offers a solution (right). The user describes the ideal person to be found — 40 year old, 80,000 Euro salary. Then the nearest neighbor algorithm starts with selecting all persons that match perfectly (again none in this case). Next, all criteria are loosened simultaneously, until the requested amount of people is found. So a nearest neighbor match engine can be seen as a search engine like AltaVista, but then for structured database information. In reality this is often implemented by calculating the distance of the relevant properties for each record to the query point, in this case the ‘ideal person’.

The nearest neighbor algorithm offers several additional information elements. The distance from a person (large dot) to the search profile (small dot) gives an indication of the relative match. A user can attach weights to specify the relative importance of the criteria. The search circle will then change into an ellipse, if the weights are not equal. For variables that cannot be expressed on a number scale, adapted distance measures can be used.

The economic approach: multi attribute utility theory

Data miners, machine learners and computer scientists often use this idea of distance based nearest neighbor search to match questions to answers or supply to demand. It is not surprising, however, that economists have developed their own concept to model these problems — although phrased differently, the so-called ‘multi attribute utility theory’ offers a very similar solution.

Assume for instance that we are looking for a second hand car, around 5,000 Euro, preferably red, and blue or green would be second best. For each criterion we can construct a so-called utility function that assigns a utility to all possible values for the criterion. In our example the utility with respect to price would be 1 for a car of 5,000 Euro, and the utility would decrease for higher or lower priced cars. The utility with respect to color would be 1 for red, 0,75 for blue or green and 0 for other colors for instance. The utility for a given second hand car would be computed by taking a weighted average of these utilities, where the weights would reflect the relative importance of the different criteria price and color. Then instead of choosing the option with the minimal distance to the optimal car (nearest neighbor) we select the car with the highest utility.


MuziekWeb voor Bezoekers - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History
Address http://www.muziekweb.nl/Bezoekers/Default.asp
Go
MuziekWeb Advies
informatie
catalogus populair
catalogus klassiek
geluidseffecten
muziekadvies
muziekstijlen
muziekwebwijzer
reacties
persoonlijke toegang
gratis
vacature
Advies laten zien

Stelt u prijs op een muziekadvies van de Centrale Discotheek? Voer dan onderstaand drie van uw favoriete artiesten in. Op basis van analyse van het leengedrag van haar leners kan de Centrale Discotheek u vervolgens een muziekadvies geven.
Voorlopig alleen populair, aan klassiek wordt gewerkt.

Artiest 1 Moby
Artiest 2 Prodigy, The
Artiest 3 Urban Dance Squad

Uitleg
Voor een muziekadvies dient u drie artiestnamen in te voeren. De schrijfwijze is **Achternaam, Voornaam** en na de komma komt een spatie. Het is niet nodig om de complete naam van de artiest in te voeren, u krijgt dan een lijst gepresenteerd waaruit u kan kiezen. Bijvoorbeeld: als u Jackson intikt krijgt u alle artiesten te zien met deze achternaam. Wanneer de drie artiesten goed zijn ingevoerd, verandert de knop "*Artiesten zoeken*" in "*Advies laten zien*".

- **Moby**
- **Prodigy, The**
- **Urban Dance Squad**

Op basis van de drie door U opgegeven artiesten geeft de Centrale Discotheek U deze suggesties. (help)

Score Artiest
Chemical Brothers, The
Faithless
Underworld
Skunk Anansie
Fatboy Slim
Blur
Massive Attack
Beastie Boys
Garbage
Jamiroquai
Portishead
Morissette, Alanis
Fun Lovin' Criminals

Figure 2
Music E-Market (http://www.muziekweb.nl): searching for music by matching with customers with similar preferences.

Matching when there are no product attributes to search on: collaborative filtering

For the sake of the argument, let us stick to the problem of matching demand and supply. There are also markets where matching on a single criteria set of product attributes (color, price, etc.) does not work out. Take for example a home improvement store that offers hundreds types of products, from nails to hammers. A nail is described with different attributes from a hammer, so it requires different matching criteria. Even if the market is constrained to one type of product, only searching on product attributes can make limited sense. On a book or CD market it only makes sense to search on 'title', 'author', or maybe 'genre'; searching on 'number of pages' or 'total playing time' makes much less sense.

A solution to this problem is not to match on product attributes but on purchase histories (also: 'collaborative filtering' or 'recommending'). A famous example is the Amazon.com personal recommendation list. But the example of a music library in Figure 2 may provide a better insight into how recommending works [Putten, 2001]. A customer can list three favorite artists. The recommender engine starts looking for customers that prefer the same artists and suggests other artists that are common in this customer target group. The counter-intuitive point of this approach is that the engine recommends artists (or albums or songs) without knowing any product attributes such as genre! In other words, the computer can make recommendations with no other information than which customers borrowed which artists.

In the remainder of this chapter we will present two cases to illustrate the data mining approach to the matching problem, with an emphasis on nearest neighbor solutions: catching criminals and hunting for jobs on an electronic marketplace.

Matching case 1: Tracking down suspects

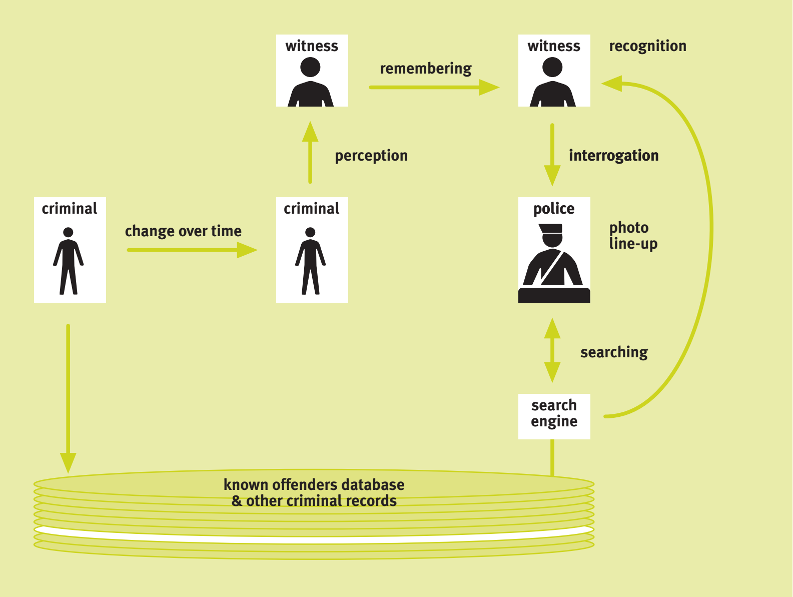
This first example comes from the government sector, more specifically law enforcement. Just like marketeers the police and other intelligence agencies are very much interested in tracking down their 'customers': criminals, terrorists, suspects and victims. It goes without saying that knowledge discovery and matching are core processes for the police and the intelligence community: hopefully investigation results in discovery. On the other hand, forensic data is often complex, polluted, fragmented, error-prone and incomplete. So data mining technologies like nearest neighbor matching can be valuable tools to search and sift through police data.

Business problem

The specific goal in this case is to find known offenders in police data given a description by a witness. The witness is then confronted with a collection of portraits from potential suspects for the purpose of identification. This called a photo line-up or a photo confrontation. Actually, this is a complex multistage process with a high number of bottlenecks and potential sources of error, shown in Figure 3.

The process starts with registration of a suspect or criminal in the police database. Research has shown that there exists substantial difference of opinion between people about descriptions. What is a skinny posture? Where does dark blond hair end and brown hair start? Is this a pale complexion or just normal? Possibly years later a crime is committed and the suspect is seen by the witness. The suspect might have changed over the years, for example he has grown bald, but added a beard or turned from adolescent into a grown-up. Furthermore, perception is an error-prone and subjective process. For example, kids are generally good observers. Characteristics like age and height, however, are likely overestimated. Or the crime has been committed outside in the dark; the witness was stressed and had little time to see the suspect.

Figure 3
The identification process.



Then, with some delay, the witness is interviewed by a police officer; a selection of potential suspects is made and shown to the witness for identification. Again, the memory of the suspect can be distorted, especially after having seen a lot of portraits.

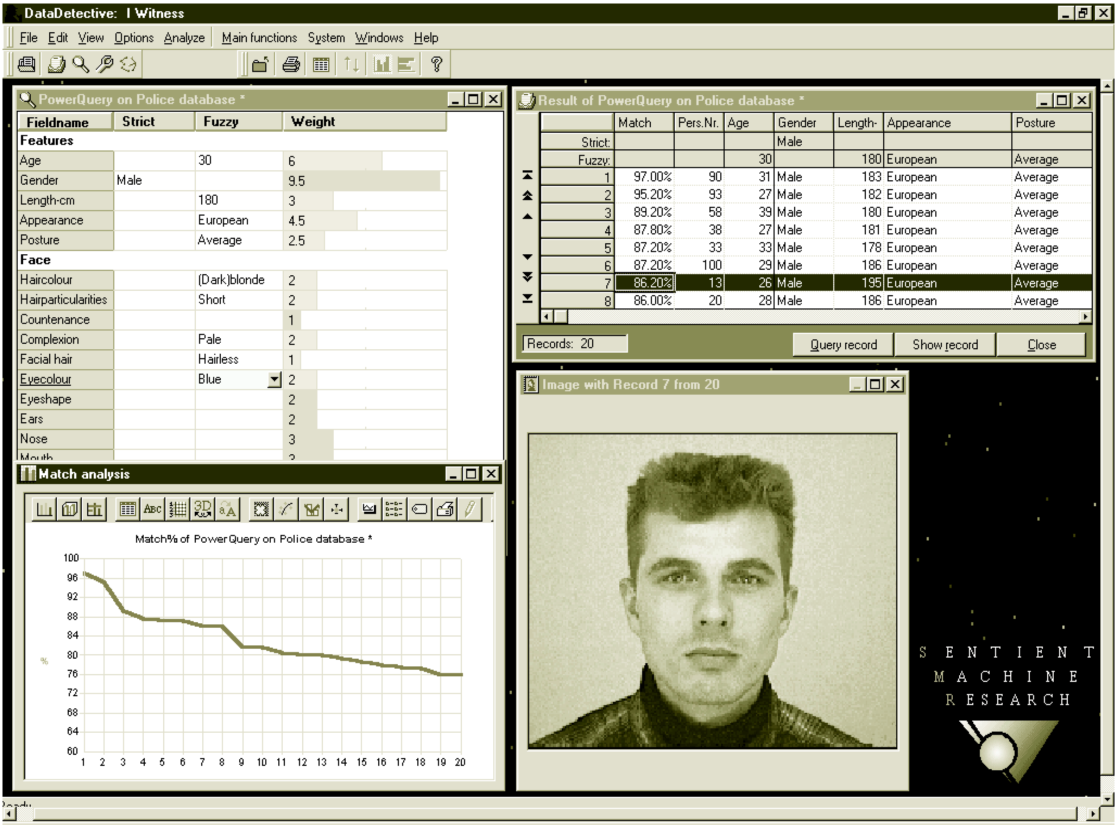
So the police officer that is responsible for making the selection is facing a dilemma. On one side he wants to use all information that the witness has given, which would result in a rich description. On the other side he likes to minimize the risk that the actual suspect is not found, because of a mismatch between the description of the witness and the registration of the known offenders in the database. It is reasonable to assume that the use of standard database queries leads to suboptimal results.

Datamining solution

Nearest neighbor matching provides a way out. The police officer uses all information available and the match engine searches for a preset number of best matching suspects.

An example of such a tool is the DataDetective Associative Recognition System[2] (Figure 4). It was based on the DataDetective Matching & Mining Engine and developed for Dutch police. In the upper left corner the police officer has specified the search query, including weights to express the relative importance or reliability of the individual criteria. In the upper right corner the resulting list of best matching criminals is shown, sorted on match score. In the lower right corner a so-called match analysis graph is drawn, to give an idea of the quality of

[2] AHS, Associatief Herkenningssysteem.


DataDetective: I Witness
File Edit View Options Analyze Main functions System Windows Help
PowerQuery on Police database *
Match analysis
Match% of PowerQuery on Police database *
100
96
92
88
84
80
76
72
68
64
60
%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Result of PowerQuery on Police database *
Records: 20
Query record
Show record
Close
Image with Record 7 from 20
SENTIENT
MACHINE
RESEARCH
Ready

Fieldname	Strict	Fuzzy	Weight
Features			
Age		30	6
Gender	Male		9.5
Length-cm		180	3
Appearance		European	4.5
Posture		Average	2.5
Face			
Haircolour		(Dark)blonde	2
Hairparticularities		Short	2
Countenance			1
Complexion		Pale	2
Facial hair		Hairless	1
Eyecolour		Blue	2
Eyeshape			2
Ears			2
Nose			3
Mouth			[illegible]

	Match	Pers.Nr.	Age	Gender	Length-	Appearance	Posture
Strict:				Male			
Fuzzy:			30		180	European	Average
1	97.00%	90	31	Male	183	European	Average
2	95.20%	93	27	Male	182	European	Average
3	89.20%	58	39	Male	180	European	Average
4	87.80%	38	27	Male	181	European	Average
5	87.20%	33	33	Male	178	European	Average
6	87.20%	100	29	Male	186	European	Average
7	86.20%	13	26	Male	195	European	Average
8	86.00%	20	28	Male	186	European	Average

Figure 4
DataDetective Associative Recognition System. Source: Sentient Machine Research.

the selection. It shows the match scores of the best matching criminals found, sorted by match score. So the graph will either stay constant or decrease steadily. From the graph and in the result list can be concluded that no one matched for 100%. In other words, if a standard database query was used, no records would have been found. In this (fictitious) search action the real suspect that was positively identified was suspect number 7 (portrait right corner below), although he did not match for the full 100%.

The matching tool is also used for photo line-ups with a slightly different purpose, the so-called evidence confrontation. In this case the police already suspect someone of having committed the crime. To make the identification task hard for the witness and fairer for the suspect the other portraits in the line-up must resemble the suspect as much as possible. The identification of one white suspect among eleven dark people would not count as evidence in court. By matching on the entire description of the suspect, high quality selections can be made quickly.

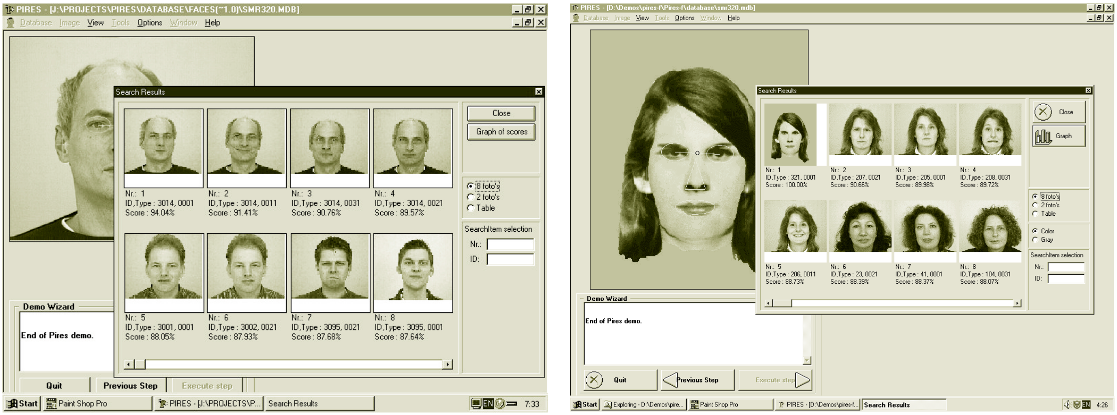
The system was evaluated by the police and the following benefits were identified:

Improvement in hit rate. A comprehensive field study was carried out in which crimes were staged and over 150 suspects were interviewed. In this case searching with the matching tool improved the hit rate by 50% compared to performance achieved with conventional query-based selections, from 22% to 34%.
Flexibility. Standard database query constrains the user to use a small number of properties that do not include errors. In an associative search system every bit of information can be specified, without worrying whether the database contains records that match exactly. The more information is used, the better the result.
Ordering by similarity. This is important, because incorrect positive identifications tend to become more frequent, if the witness has already seen more pictures. So it is important to show the best candidates early in the selection.
User friendliness. Previously, selection was the work of specialists and creating a selection of adequate size took a lot of time. With this tool users without special training or experience can search the database; there is no need to develop a strategy to cope with the rigidity of standard database query systems. Actually, the field study showed that users without special photo confrontation training performed better than the specialists! This means that a lot more photo confrontations can be carried out, improving the absolute number of hits even more.

Future prospects

In a police investigation, police officers are constantly trying to match clues to all kinds of known information. Matching tools may be developed that can perform a holistic match on a wide variety of data simultaneously. This might include all kinds of relational and multimedia data such as information on cases, cars, fingerprints, shoe tracks, faces, phone, e-mail, documents, GPS data, DNA and so on. Of course, legal aspects regarding privacy and law enforcement regulations are major issues here.

An example is face recognition (the Pires Face Search & Description System, Figure 5). The tool recognizes a face in a portrait, generates a description and searches for similar ones. The technology behind it is based on a large collection of neural networks and an expert system to generate descriptions and a match engine to match on them. Pires can be used during registration, to generate a description advice and to double-check whether the person registered is already known to the police. The tool can also aid investigation or observation purposes where identification is needed. For instance, it is possible to use a photo composition picture as input. Pictures of unidentified murder victims could also be entered in Pires to find out whether the police know the victim. Finally, Pires could be used in evidence confrontations to select persons similar to the suspect to be identified. This is to make sure that a confrontation procedure is relatively hard for the witness and thus fair.


PIRES - [J:\PROJECTS\PIRES\DATABASE\FACES(~1.0)\SMR320.MDB]

Search Results

Nr.: 1
ID,Type : 3014, 0001
Score : 94.04%

Nr.: 2
ID,Type : 3014, 0011
Score : 91.41%

Nr.: 3
ID,Type : 3014, 0031
Score : 90.76%

Nr.: 4
ID,Type : 3014, 0021
Score : 89.57%

Nr.: 5
ID,Type : 3001, 0001
Score : 88.05%

Nr.: 6
ID,Type : 3002, 0021
Score : 87.93%

Nr.: 7
ID,Type : 3095, 0021
Score : 87.68%

Nr.: 8
ID,Type : 3095, 0001
Score : 87.64%

● 8 foto's
○ 2 foto's
○ Table

SearchItem selection
Nr.:
ID:

Demo Wizard
End of Pires demo.

PIRES - [D:\Demos\pires-f\Pires-f\database\smr320.mdb]

Search Results

Nr.: 1
ID,Type : 321, 0001
Score : 100.00%

Nr.: 2
ID,Type : 207, 0021
Score : 90.66%

Nr.: 3
ID,Type : 205, 0001
Score : 89.98%

Nr.: 4
ID,Type : 208, 0031
Score : 89.72%

Nr.: 5
ID,Type : 206, 0011
Score : 88.73%

Nr.: 6
ID,Type : 23, 0021
Score : 88.39%

Nr.: 7
ID,Type : 41, 0001
Score : 88.37%

Nr.: 8
ID,Type : 104, 0031
Score : 88.07%

● 8 foto's
○ 2 foto's
○ Table

● Color
○ Gray

SearchItem selection
Nr.:
ID:

Demo Wizard
End of Pires demo.

Database Image View Tools Options Window Help
Close
Graph of scores
Quit
Previous Step
Execute step
Start
Paint Shop Pro
PIRES - [J:\PROJECTS\P...
Search Results
7:33
Close
Graph
Exploring - D:\Demos\pire...
PIRES - [D:\Demos\pires-f...
4:26

Figure 5

Pires system. Searching for similar portraits using face recognition and matching. Input can be a photo (left) or a composition portrait (right).

In future forensic information systems we will see more and more ‘like-this’ functionality in the sense that wherever a user navigates through the various databases in his intranet environment, even if he is using a standard database or keyword search, he will always be able to pull up more cases like this, persons like this, crime scenes like this, etc. You can imagine that a small collection of matching cases and persons is always active in some frame of his screen depending on the context, thus allowing for serendipitous discoveries to happen more often.

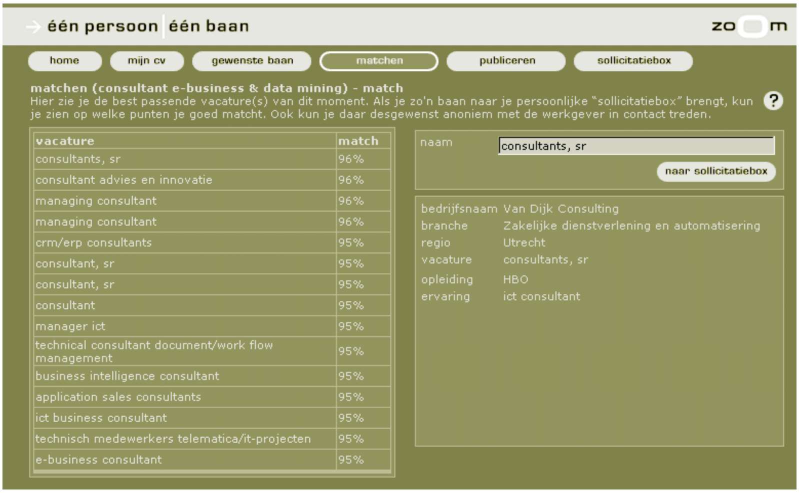
Matching case 2: Job matching

Another typical application of matching technology is the job market. The case will describe job matching and mining services offered by Matchcare [Putten, 1999].

Business problem

By describing the job matching case we intend to provide a relevant and frequently occurring example of an E-market. On one side, every organization deals with the selection and retention of personnel. On the other side, people are changing jobs frequently nowadays and are becoming more and more responsible for managing their own careers. Obviously, the friction between supply and demand is an important problem for individuals, organizations and the economy as a whole. A variety of intermediary players try to bridge the gap. For example, government agencies start projects to get the unemployed back to work, headhunters scout for high potential personnel, human resource departments provide job rotation intranets and television, print media and the web are flooded with job adverts.

In reality, the ideal job may not exist (nor does the ideal applicant). And even if it did exist, it might not have been known to the potential employee (or the employer). The Dutch company Matchcare addresses these problems by offering job seekers, employers and intermediaries on the job market solutions



matchen (consultant e-business & data mining) - match

Hier zie je de best passende vacature(s) van dit moment. Als je zo'n baan naar je persoonlijke "sollicitatiebox" brengt, kun je zien op welke punten je goed matcht. Ook kun je daar desgewenst anoniem met de werkgever in contact treden.

vacature	match
consultants, sr	96%
consultant advies en innovatie	96%
managing consultant	96%
managing consultant	96%
crm/erp consultants	95%
consultant, sr	95%
consultant, sr	95%
consultant	95%
manager ict	95%
technical consultant document/work flow management	95%
business intelligence consultant	95%
application sales consultants	95%
ict business consultant	95%
technisch medewerkers telematica/it-projecten	95%
e-business consultant	95%

naam: consultants, sr

bedrijfsnaam	Van Dijk Consulting
branche	Zakelijke dienstverlening en automatisering
regio	Utrecht
vacature	consultants, sr
opleiding	HBO
ervaring	ict consultant

Figure 6

Job market example application. Selection of best matching jobs.

based on web, data mining and matching technology. The main technical challenge is to provide a model that offers enough flexibility to develop solutions for different target groups, but with low costs and the benefits of reusing information and core matching and mining technologies.

Datamining solution

The solution model consists of a small number of core components. All content — jobs and resumes — is centrally stored using a uniform, standardized data model that captures the essence of a job or resume (the 'ontology'). A central match engine provides matching services on subsets of the database. Data mining services are offered to provide knowledge derived from this data. The system can be exploited through various channels and applications e.g. target group specific web sites for the general public, university students , unemployed people who take part in a regional reintegration program, etc.

An example service that has been developed within this model was a job market containing vacancies that appear in Dutch print media.
These vacancies are scanned from magazines and newspapers. Then over two hundred properties may be derived from the vacancies, such as profession offered, industrial sector, type of work, required skills and education, salary range, benefits, contact information, etc. The derivation of these properties is part manual, but also part automated by carrying out text mining on the vacancy and the history of previous vacancies. Some properties are derived using background information such as classifications of industrial sectors and professions that are provided by various statistical and research agencies.
On the web site, consumers may look for jobs for free. Employers may see CV's of job seekers, but only with explicit permission and they have to pay a fee. Job seekers may fill in their resumes with information on education and experience,

build different profiles for each of the target jobs they would like to have, publish their profile, run a match search on the database and apply on-line.

In Figure 6 an example is given. The match engine performs a nearest neighbor search on the resumes and the target job profile. In this case the target job is 'consultant data mining and E-business' and the result includes jobs like consultants for innovation, customer relationship management, business intelligence and E-business. It is possible to apply on-line without giving up anonymity. Before applying, a high-level match analysis can be run to determine the match on the different parts of the profile. This way an applicant gets an idea about stronger and weaker points before the interview. In this example the match was 100% for all that the applicant had to offer: skills, experience and education. However, there was a small mismatch (3%) between the ideal target job and the job offered.

The same job content and core services are used to aid various government and commercial organizations in projects to reintegrate people who are unemployed, for instance because of partial disabilities. In one such a project clients not only receive matching vacancies, but they also receive very detailed information on their strength and weaknesses and suggestions for improvement, given the vacancies they match best for. Data mining techniques such as deviation detection and profiling algorithms can be used for this, in addition to nearest neighbor search.

Vision for the future

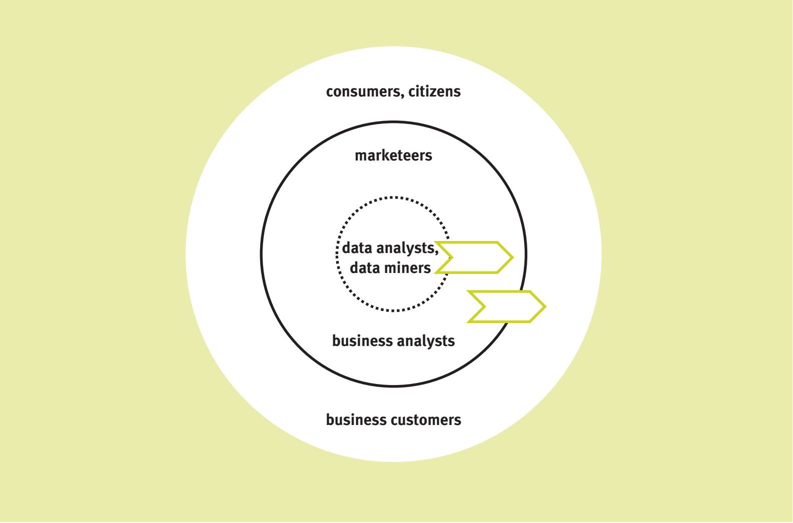
So, what can be expected with respect to matching in the future?
First note that in both case examples the actual users of the data mining system were not data miners at all, but rather policemen or site visitors who are probably unaware of the term data mining in the first place. This ties in with an important trend: the democratization of data mining (Figure 7). In the early days, data mining was an apparently magic technology that could only be used by scientists, quantitative or statistical analysts or data miners. However, since 1995, data mining started to become more accessible for data analysis savvy marketeers and business analysts. In future more and more consumers, citizens and business customers will be able to profit from data mining, end users that don't know much (if anything) about data mining implementations or technologies. From a technical data miner's point of view this may not seem important or maybe even not desirable — but just compare the number of data mining analysts with for instance the number of Amazon.com customers and it will be clear what profound effect it will have on the value and viability of data mining in future. Typically, matching will be on the forefront of this development, because it is a more end-user focused data mining task than clustering or prediction for instance.

If we constrain ourselves to match market — product search applications, there have been some interesting developments over the past years. The first applications were geared towards shop bots for business to consumer sites: think shops and portals for jobs, cars, houses, boats, etc. Then business to business E-match markets (also known as net markets) became en vogue, linking buyers and sellers within a specific vertical: chemicals, plastics, automotive, metals, agro-food, etc. Now businesses become more aware that procurement is actually a strategic activity that shouldn't be shared with competitors, and a new industry is coming up that provides private match marketplaces and auctions for a single buyer: strategic sourcing (E-sourcing) marketplaces. In future there will be room for all three of these application types, each with it's ups and downs. For instance, since the dot.com burst the interest in business to consumer applications has waned, but specialized companies building products such as recommendation systems for video hard disks (Tivo, etc.) are being founded and funded.

There are actually a lot of fruitful cross links to other data mining tasks for matching applications and markets. Prediction for instance can be used for negotiation support: what is a decent salary for me given the jobs I match best with? Is this car over- or underpriced? Clustering and dimension reduction can be used to project the high dimensional space of product attributes into two or three dimensions, thus allowing users to navigate themselves through the space of houses offered for instance, and find the best match.

All of these trends will work towards the same vision: more and more people will be data mining every day — without even noticing it.

Figure 7
Data mining democratization.



References

- Guttman, R., P. Maes. (1998). Agent Mediated Integrative Negotiation for Retail Electronic Commerce. Proceedings of the Workshop on Agent Mediated Electronic Trading (AMET'98), Minneapolis, Minnesota
- Putten, P. van der. (1999). Datamining in Bedrijf. Informatie en Informatiebeleid **17**:3
- Putten, P. van der, M.J. den Uyl. (2001). Mining E-markets. IT Monitor **3**. Ten Hagen & Stam