Institute of Computer Science
Technical University of Łódź, Poland

News Generating via Fuzzy Summarization of Databases

Adam Niewiadomski

aniewiadomski@ics.p.lodz.pl

The assumptions

- Huge datasets are necessary to manage and control e.g.:

  rescue and defence systems,

  weather forecasting,

  human resources,

  mass media, telecommunication, etc.

  but

- The amount of available information exceeds human perception (e.g. Google)

The assumptions

- Huge datasets are necessary to manage and control e.g.:

  rescue and defence systems,

  weather forecasting,

  human resources,

  mass media, telecommunication, etc.

  but

- The amount of available information exceeds human perception (e.g. Google)

- **Interpretation and knolwedge**, and not raw numbers, **should be provided** by IT, e.g.

  About 1/4 people of Asia suffer from hunger

  instead of

  83,617,497 people of India and 222,987,363 people of China and 3,987,256 people of Tailand and 483,987 people of Pakistan...

The scope

- In general:

  to obtain knowledge from information

- In detail:

  to generate a brief textual message from a large[a] numerical dataset

[a] Large – impossible to be processed "manually" in a reasonable time.

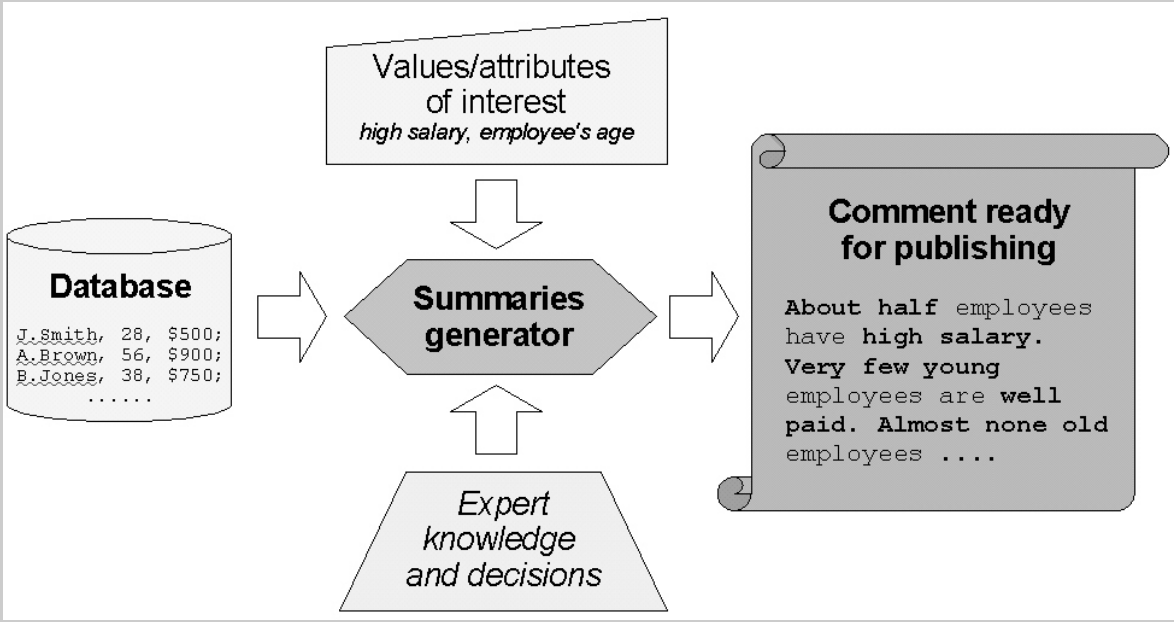
The scope

- In general:

  to obtain knowledge from information

- In detail:

  to generate a brief textual message from a large[a] numerical dataset



[a] Large – impossible to be processed "manually" in a reasonable time.

Agenda

.: Fuzzy sets – linguistic knowledge representation

.: Linguistic summaries of databases

.: Automated generating of textual news/comments

.: Implementation

.: Results and further work

Fuzzy sets (1)

- A crisp (classic) set K in a universe $\mathcal{X}$
  is represented by its characteristic function $\chi_K : \mathcal{X} \to \{0, 1\}$

$$x \in \mathcal{K} \vee x \notin K \tag{1}$$

  and *tertium non datur*.

Fuzzy sets (1)

- A crisp (classic) set K in a universe $\mathcal{X}$
  is represented by its characteristic function $\chi_K: \mathcal{X} \to \{0, 1\}$

$$x \in \mathcal{K} \vee x \notin K \tag{1}$$

  and *tertium non datur*.

- A fuzzy set A in a (non-empty) universe $\mathcal{X}$
  is represented by its membership function $\mu_A: \mathcal{X} \to [0, 1]$ (Zadeh, 1965)

$$A =_{df} \{< x, \mu_A(x) >: x \in \mathcal{X}\} \tag{2}$$

  hence x belongs to A at the grade of $\mu_A(x)$, $0 \leq \mu_A(x) \leq 1$

Fuzzy sets (2)

- Fuzzy sets can represent imprecise concepts/statements in natural languages:

  tall man, high temperature — imprecise features/properties

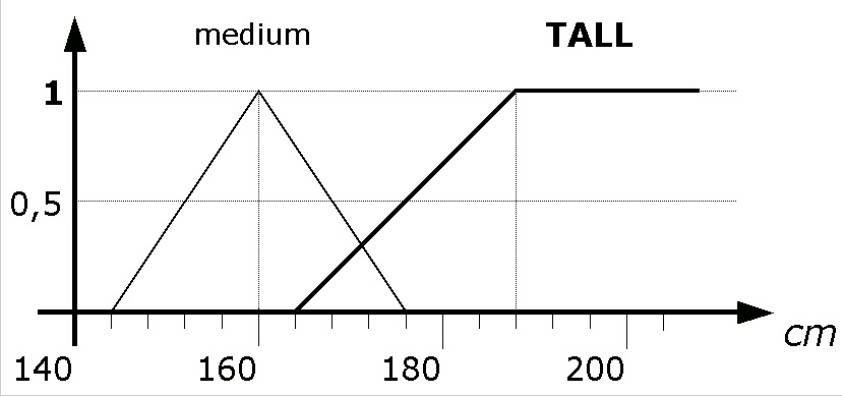
  about 1/4, **much more than 1000** — imprecise quantities

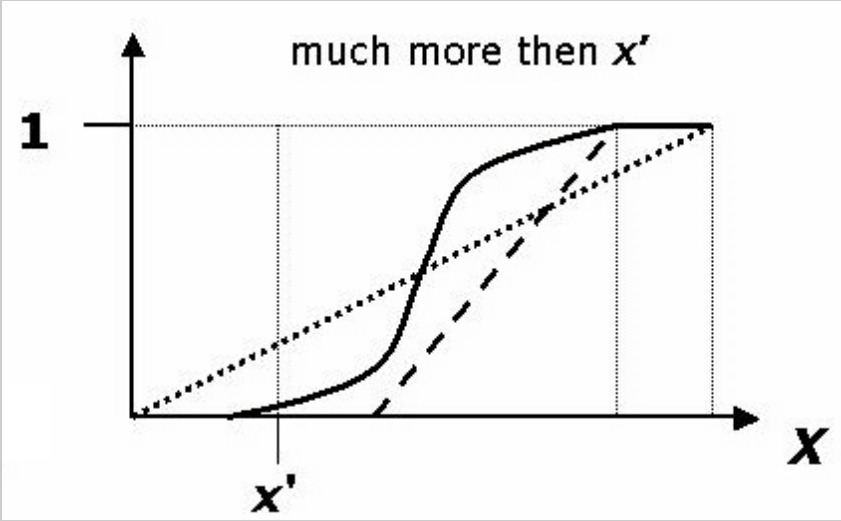
Fuzzy sets (2)

- Fuzzy sets can represent imprecise concepts/statements in natural languages:

  tall man, high temperature — imprecise features/properties

  about 1/4, **much more than 1000** — imprecise quantities





A linguistic summary of a database... (1)

...is a semi-natural sentence (Yager, 1982)

$$Q\ P \text{ are/have } S\ [T] \tag{3}$$

Q – a quantity in agreement, fuzzy quantifier, e.g. *about half*

P – a subject of the summary, e.g. *cars, workers*

S – a summarizer, a property of interest, e.g. *fast, young*

$T \in [0, 1]$ – the degree of truth of the summary

A linguistic summary of a database... (1)

...is a semi-natural sentence (Yager, 1982)

$$Q\ P \text{ are/have } S\ [T] \tag{3}$$

Q – a quantity in agreement, fuzzy quantifier, e.g. *about half*

P – a subject of the summary, e.g. *cars, workers*

S – a summarizer, a property of interest, e.g. *fast, young*

$T \in [0, 1]$ – the degree of truth of the summary

Q and S are represented by fuzzy sets, hence

$$T = \mu_Q\left(\frac{\sum_{i=1}^{m}\mu_S(d_i)}{m}\right) \tag{4}$$

e.g.

$$\underbrace{\textit{About half}}_{Q}\ \underbrace{\text{workers}}_{P}\ \text{are}\ \underbrace{\textit{young}}_{S}\ \underbrace{[0.62]}_{T} \tag{5}$$

Linguistic summaries of databases (2)

- A composite summarizer (George, Srikanth, 1996)

$$S = S_1 \text{ and } S_2 \text{ and } \ldots \text{ and } S_n \tag{6}$$

e.g. Very few workers are *young* and *well-paid*

Linguistic summaries of databases (2)

- A composite summarizer (George, Srikanth, 1996)

$$S = S_1 \text{ and } S_2 \text{ and } \ldots \text{ and } S_n \tag{6}$$

  e.g. Very few workers are *young* and *well-paid*

- A summary with a *query* (Kacprzyk, Yager, 2001)

$$Q\ P \text{ being } w_g \text{ are/have } S\ [T] \tag{7}$$

  e.g. Many young girls are pretty

Linguistic summaries of databases (2)

- A composite summarizer (George, Srikanth, 1996)

$$S = S_1 \text{ and } S_2 \text{ and } \ldots \text{ and } S_n \tag{6}$$

  e.g. Very few workers are *young* and *well-paid*

- A summary with a *query* (Kacprzyk, Yager, 2001)

$$Q\ P \text{ being } w_g \text{ are/have } S\ [T] \tag{7}$$

  e.g. Many young girls are pretty

- Other measures of quality of summaries (Kacprzyk, Yager, Zadrożny, 2001)

  imprecision, covering, appropriateness, length of the summary, etc.

News and comments generating

- Database

$$\begin{aligned}\mathcal{D} &= \{ \quad < V_1(y_1), V_2(y_1), ..., V_n(y_1) >, \dots \\ &\dots \quad < V_1(y_m), V_2(y_m), ..., V_n(y_m) > \} = \{d_1, d_2, ..., d_m\}\end{aligned}$$

y_i – a real object (e.g. car, man);

V_j – an attribute (e.g. speed, age);

d_i – the record describing y_i (e.g. $<$ Ford, 1999, 200 km/h$>$)

News and comments generating

- Database

$$\begin{aligned}\mathcal{D} &= \{ \quad < V_1(y_1), V_2(y_1), ..., V_n(y_1) >, \dots \\ &\dots \quad < V_1(y_m), V_2(y_m), ..., V_n(y_m) > \} = \{d_1, d_2, ..., d_m\}\end{aligned}$$

y_i – a real object (e.g. car, man);

V_j – an attribute (e.g. speed, age);

d_i – the record describing y_i (e.g. $<$ Ford, 1999, 200 km/h$>$)

- The features of interest – summarizers represented by fuzzy sets

S_1=low, S_2=medium, S_3=more than 130 for V_1=SPEED

S_4=young, S_5=about 30, S_6=old for V_2=AGE, etc.

News and comments generating

- Database

$$\begin{aligned}\mathcal{D} &= \{ < V_1(y_1), V_2(y_1), ..., V_n(y_1) >, \dots \\ &\dots < V_1(y_m), V_2(y_m), ..., V_n(y_m) > \} = \{d_1, d_2, ..., d_m\}\end{aligned}$$

y_i – a real object (e.g. car, man);
V_j – an attribute (e.g. speed, age);
d_i – the record describing y_i (e.g. $<$ Ford, 1999, 200 km/h$>$)

- The features of interest – summarizers represented by fuzzy sets

S_1=low, S_2=medium, S_3=more than 130 for V_1=SPEED
S_4=young, S_5=about 30, S_6=old for V_2=AGE, etc.

- Linguistic quantifiers represented by fuzzy sets in $\mathbb{R} \cup \{0\}$

Q_1=few, Q_2=about half, Q_3=not less than 500, etc.

Algorithms

1. for each single summarizer $S \in \{S_1, ..., S_z\}$ // generating (Y) summaries

   1.1. for each quantifier Q_h, $h = 1, ..., k$

   if (Q_h is absolute)

   compute $T_h = \mu_{Q_h}\left(\sum_{j=1}^{m} \mu_S(d_j)\right)$

   else // i.e. if Q_h is relative

   compute $T_h = \mu_{Q_h}\left(\frac{\sum_{j=1}^{m} \mu_S(d_j)}{m}\right)$

   1.2. compute $T_{h_{\max}} = \max_{h=1,...,k} T_h$, remember $h_{\max}$

   1.3. generate summary in the form of

   $Q_{h_{\max}}$ P is/have S $[T_{h_{\max}}]$

.....

Algorithms

1. for each single summarizer $S \in \{S_1, ..., S_z\}$ // generating (Y) summaries

   1.1. for each quantifier Q_h, $h = 1, ..., k$

   if (Q_h is absolute)

   compute $T_h = \mu_{Q_h}\left(\sum_{j=1}^{m} \mu_S(d_j)\right)$

   else // i.e. if Q_h is relative

   compute $T_h = \mu_{Q_h}\left(\frac{\sum_{j=1}^{m} \mu_S(d_j)}{m}\right)$

   1.2. compute $T_{h_{\max}} = \max_{h=1,...,k} T_h$, remember $h_{\max}$

   1.3. generate summary in the form of

   $Q_{h_{\max}}$ P is/have S [$T_{h_{\max}}$]

.....

The number of all possible summaries is

$$k \sum_{i=0}^{z-1} \binom{z}{i} (2^{z-i} - 1) \tag{8}$$

Implementation

Acknowledgements for the graduating student of mine – **Mr. Marcin Białas**

- Database – MS Access (*.mdb) or MS SQL Server (*.mdf), ca 10,000 records
- GUI and fuzzy logic engine – C# and .NET Framework 1.1

Implementation

Acknowledgements for the graduating student of mine – **Mr. Marcin Białas**

- Database – MS Access (*.mdb) or MS SQL Server (*.mdf), ca 10,000 records
- GUI and fuzzy logic engine – C# and .NET Framework 1.1
- Two software components:

  Manager – expert knowledge acquisition

  Generator – producing news

Further work directions

- In theory:

  To formalize Type-2 Linguistic Summaries of Databases, since

  $$\text{Fuzzy Sets} \subset \text{Interval-Valued Fuzzy Sets} \subset \text{Type-2 Fuzzy Sets}$$

  To find and test new quality indices of summaries (increasing their objectivity)

Further work directions

- In theory:

  To formalize Type-2 Linguistic Summaries of Databases, since

  $$\text{Fuzzy Sets} \subset \text{Interval-Valued Fuzzy Sets} \subset \text{Type-2 Fuzzy Sets}$$

  To find and test new quality indices of summaries (increasing their objectivity)

- In practice:

  To improve linguistic performance (correctness, grammar) of generated messages (especially in slavonic languages – czech, slovak, polish, etc.)

  To improve GUI, e.g. visualisation and freehand drawing of membership functions

Literature references

1. Niewiadomski, A., Ochelska, J., Szczepaniak, P. S., Interval-valued linguistic summaries of databases, *Control and Cybernetics*, Vol. 2, 2005, (in print)

2. Niewiadomski, A., On Two Possible Roles of Type-2 Fuzzy Sets in Linguistic Summaries, *Lecture Notes in Artificial Intelligence*, Vol. 3528, 2005, pp. 341-347.

3. Niewiadomski, A., Interval-valued linguistic variables. An application to linguistic summaries, *Issues in Intelligent Systems*. Paradigms. In: O. Hryniewicz, J. Kacprzyk, J. Koronacki, S. T. Wierzchon (Eds.), EXIT Academic Press, Warsaw, 2005, pp. 167-183.

4. Niewiadomski, A., Interval-valued quality measures for linguistic summaries, *Issues in Soft Computing. Theory and Applications*. In: P. Grzegorzewski, M. Krawczak, S. Zadrozny (Eds.), EXIT Academic Press, Warsaw, 2005, pp.211–224

5. Niewiadomski, A. Bartyzel, M., Elements of Type-2 Semantics in Linguistic Summaries of Databases, *Lecture Notes in Artificial Intelligence*, 2006, (submitted).

6. Niewiadomski, A. Szczepaniak, P. S., News generating based on Interval Type-2 Linguistic Summaries of Databases, *Proceedings of IPMU 2006 Conference*, July 2-7, 2006, Paris, France, 2006, (submitted).

Thank you very much for your kind attention