Evaluating Software Project Similarity by using Linguistic Quantifier Guided Aggregations

Ali IDRI, ENSIAS, Rabat, Morocco
Alain ABRAN, UQAM, Montreal, Canada

Joint 9th IFSA World Congress and
20th NAFIPS International Conference

25-28 July, Vancouver, Canada 2001

UQÀM

Summary

- Software Project Similarity Measures: Background and Related Work
- Linguistic Quantifiers
- Limits of the Existing Similarity Measures
- Improvements by Using Linguistic Quantifiers
- Illustration with COCOMO'81 Dataset
- Conclusions and Future Work

Software Project similarity Measures

- **Software Project similarity is one of the most important process software attribute**
- **It is often used when estimating software development effort by analogy**
- **Intuitively, two software projects are not similar if the <u>differences</u> between their <u>sets of attributes</u> are <u>obvious</u>**
- **Analogy**

Human	Software Project
Size	Size
Color of skin	Complexity
Color of eyes	Reliability
Color of hair	Analysts competency
...	...

Related Work

Sheppered et al. (1997)

$$d(P_1,P_2,V) = \frac{1}{\sum_{v_j} d_{v_j}(P_1,P_2)} \quad d_{v_j}(P_1,P_2) = \begin{cases} (v_j(P_1) - v_j(P_2))^2 \\ \begin{cases} 0 & if \quad v_j(P_1) = v_j(P_2) \\ 1 & if \quad v_j(P_1) \neq v_j(P_2) \end{cases} \end{cases}$$

Idri and Abran, 7th FT&T, Atlantic City, 2000

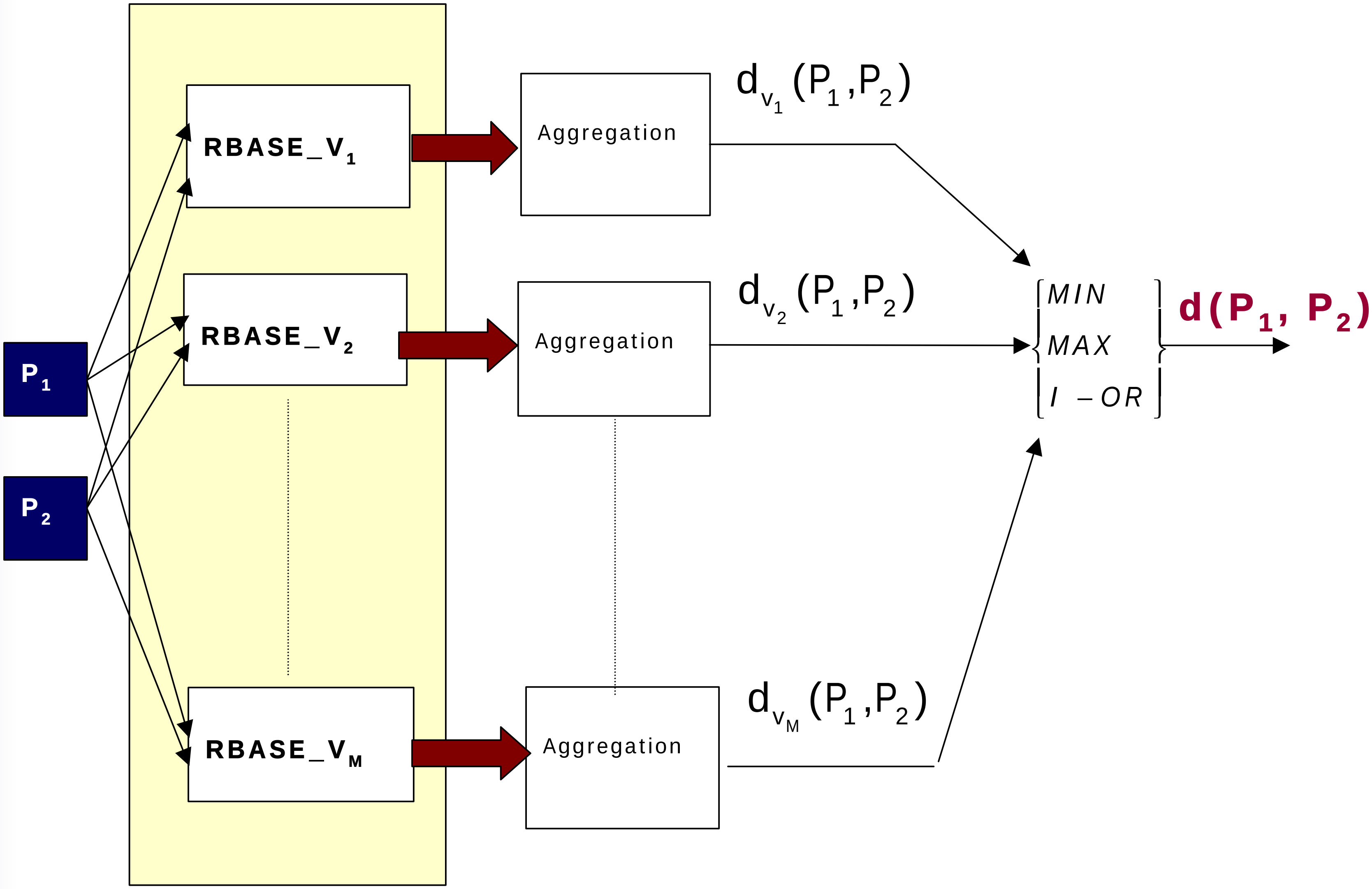
The equality distance is not precise and can give great difference when estimating effort for two similar software projects described by **Vagueness information**

Idri and Abran, 6th MCSEAI, Morocco, 2000

We have proposed a set of similarity measures based on fuzzy logic

UQÀM

RBASE


RBASE_V1
RBASE_V2
RBASE_VM
P1
P2
Aggregation
Aggregation
Aggregation
$d_{V_1}(P_1,P_2)$
$d_{V_2}(P_1,P_2)$
$d_{V_M}(P_1,P_2)$
$\begin{Bmatrix} MIN \\ MAX \\ I-OR \end{Bmatrix}$
$d(P_1, P_2)$

Idri and Abran, 7th IEEE Metrics, London, 2001

After an axiomatic validation, we have retained the following similarity measures :

$$d_{v_j}(P_1,P_2) = \mu_{\underset{\approx}{R}^{v_j}}(P_1,P_2) = \begin{cases} \max_k \min(\mu_{A_k}^{v_j}(P_1), \mu_{A_k}^{v_j}(P_2)) \\ \max-\min \;\; aggregation \\ ou \\ \sum_k \mu_{A_k}^{v_j}(P_1) \times \mu_{A_k}^{v_j}(P_2) \\ sum - product \;\; aggregation \end{cases}$$

$$d(P_1,P_2) = \begin{cases} \min(d_{v_1}(P_1,P_2),\ldots,d_{v_M}(P_1,P_2)) \\ \max(d_{v_1}(P_1,P_2),\ldots,d_{v_M}(P_1,P_2)) \\ i-or(d_{v_1}(P_1,P_2),\ldots,d_{v_M}(P_1,P_2)) = \begin{cases} 0 \quad \exists k,h / d_{v_k}(P_1,P_2) = 1 \text{ and } d_{v_h}(P_1,P_2) \\ \dfrac{\prod_{j=1}^{M} d_{v_j}(P_1,P_2)}{\prod_{j=1}^{M}(1-d_{v_j}(P_1,P_2)) + \prod_{j=1}^{M} d_{v_j}(P_1,P_2)} \quad \text{otherwise} \end{cases} \end{cases}$$

Objective

To improve our similarity measures by using a **soft aggregation** of the individuals similarities

Similarity measures will be easily calibrated and adapted to the needs and the characteristics of each organization

Linguistic quantifiers

- **Human discourse uses a large number of linguistic quantifiers**
- **Zadeh distinguishes between two classes:**
  - Absolute linguistic quantifiers
  - Proportional linguistic quantifiers (**most**, **few**, **at least**, **at most**,...)
- **Yager has distinguished three categories of proportional quantifiers:**
  - RIM quantifiers (**most**, **at least a**,...)
  - RDM quantifiers (**few**, **at most a**,...)
  - RUN quantifiers (**about a**)

UQÀM

Limits of the existing similarity measures

- **In the previous work, we have used only two RIM quantifiers '<u>all</u>' and '<u>there exists</u>' to combine the individual distances**

- **Critics:**
  - The **'all'** and the **'there exists'** quantifiers are not always a good combination:

$$d_{v_{j0}}(P_1, P_2) = 0 \text{ (or 1)}, \quad d_{v_j}(P_1, P_2) = 1 \text{ (or 0) for } j \neq j_0$$

When we use a *min* (or *max*) operator, the overall distance $d(P_1, P_2)$ is null (or equal to 1), while a suitable combination would seem to give a value in the vicinity of 1 (or 0);

- In many situations, other linguistic quantifiers can be useful such that **'most'**, **'many'**, and **'at least a'**

- We must take into account the importance of the variables describing the software projects because often the influence of some variables is greater than of others

- The **i-or** operator?

1. **It has no clear natural interpretation**

$$\frac{ab}{(1-a)(1-b)+ab} = w_1 a + w_2 b \quad with \quad \begin{cases} w_1, w_2 \in [0,1] \\ w_1 + w_2 = 1 \end{cases}$$

2. **Suppose that we have**

$$d_{v_{j0}}(P_1, P_2) = 1, \quad d_{v_j}(P_1, P_2) - \approx 0 \text{ for } j \neq j_0$$

When applying the I-or operator, the overall distance $d(P_1,P_2)$ is equal to 1, while a suitable combination seems to give a result other than 1.

UQÀM

Improvements by using Linguistic Quantifiers

Solution

Evaluation of the $d(P_m,P_n)$ by aggregating the individual distances using RIM linguistic quantifiers

$$d(P_m,P_n) = \begin{cases} all \;\; of \;\; d_{v_j}(P_m,P_n) \\ most \;\; of \;\; d_{v_j}(P_m,P_n) \\ many \;\; of \;\; d_{v_j}(P_m,P_n) \\ at \;\; least \;\; four \;\; of \;\; d_{v_j}(P_m,P_n) \\ \ldots \\ there \;\; exists \;\; of \;\; d_{v_j}(P_m,P_n) \end{cases}$$

UQÀM

RIM linguistic quantifiers is implemented by OWA operators

- We must provide the appropriate linguistic quantifier to be used in an organization, **Q**
- The linguistic quantifier, **Q**, is used to generate an OWA weighting vector **W (w_1, w_2, …, w_M)**
- We calculate the overall similarity by :

$$d(P_1, P_2) = \sum_{j=1}^{M} w_j d_{v_j}(P_1, P_2)$$

$$w_j(P_1, P_2) = Q\left(\frac{\sum_{k=1}^{j} u_k}{T}\right) - Q\left(\frac{\sum_{k=1}^{j-1} u_k}{T}\right)$$

Illustration with COCOMO'81 dataset

- **COCOMO'81 dataset is composed of 63 software projects**

- **Each project is described by 17 attributes:**

  - Software size measured in **KDSI**

  - Project Mode is defined as **Organic**, **Semi-detached** or **Embedded**

  - 15 cost drivers related to the software environment

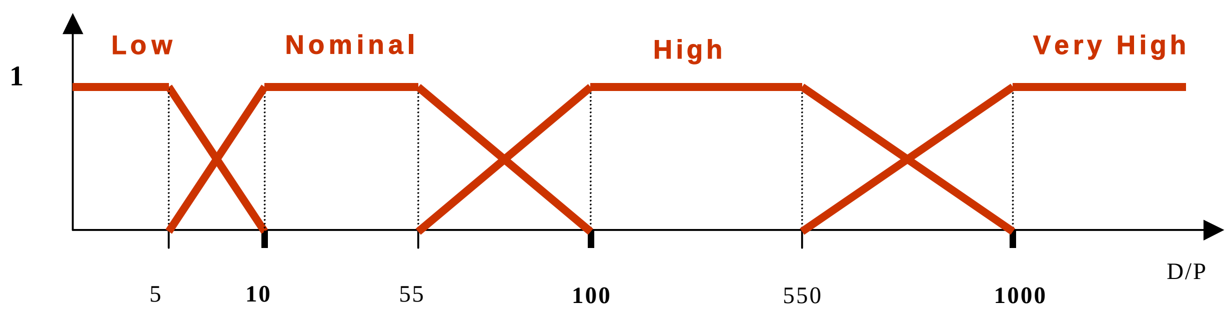
    Very low, Low, Nominal, High, Very high, Extra-high

Example: DATA cost driver

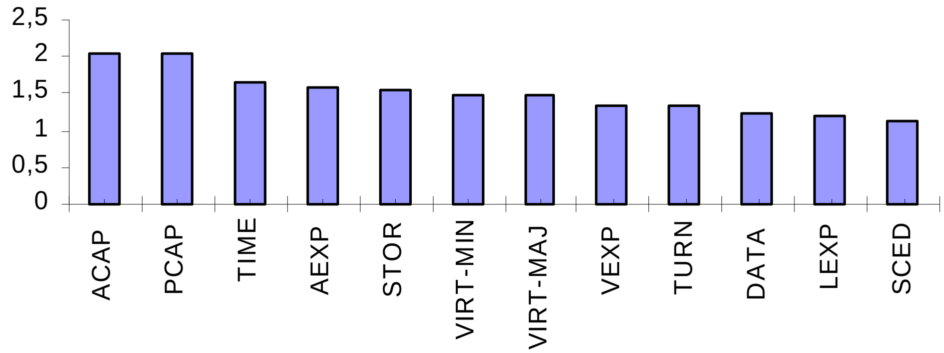
$$\frac{D}{P} = \frac{\text{Database size in bytes or characters}}{\text{Program size in DSI}}$$

Low	Nominal	High	Very high
D/P<10	10<=D/P<100	100<=D/P<1000	D/P>=1000

- It is more general
- It mimics the way in which humans interpret linguistic values
- The transition from one linguistic value to a contiguous linguistic value is **gradual** rather than **abrupt**



- For simplification purpose, we calculate only the similarity between the first project and the first five projects of the dataset
- Our measures are computationally intensive; so we have developed a software prototype with VB and MS-access
- The prototype allows us to try various RIM linguistic quantifiers **Q** to the COCOMO'81 dataset
- The weights U_k is calculated by means of the project's productivity ratio :



- **In this illustration, we use RIM linguistic quantifiers defined by :**

$$Q(r) = r^{\alpha} \quad (\alpha > 0)$$

- **We use only the max-min aggregation to calculate the individual similarities:**

  - If all the fuzzy sets associated to software project attributes are **normal**, **convex** and form a **fuzzy partition** then max-min and sum-product aggregations give approximately the same results

  - The sum-product aggregation does not respect all axioms

$$d(P_i, P) \leq d(P, P) \; ?$$

		Max-min aggregation $d_{v_j}(P_1, P_n)$					
		d(P_1, P_n)					
			P_1	**P_2**	**P_3**	**P_4**	**P_5**
P_1	**a**	Max	**1**	**1**	**1**	**1**	**0,84096**
		1/100	0.99824	0.98529	0.97938	0.99095	0.82482
		1/30	0.99416	0.95189	0.93298	0.97018	0.78841
		1/20	0.99128	0.92879	0.90124	0.95564	0.76343
		1/15	0.98842	0.90629	0.87061	0.94134	0.73927
		1/10	0.98275	0.86304	0.81253	0.91344	0.69331
		1/7	0.97559	0.81069	0.74370	0.87890	0.63855
		1/5	0.96625	0.74617	0.66117	0.83505	0.57245
		1/3	0.94533	0.61618	0.50335	0.74178	0.44446
		1	0.85691	0.24612	0.132939	0.417857	0.13026
		3	0.69875	2.0948E-02	2.9783E-03	8.4882E-02	4.4606E-03
		5	0.62337	2.3918E-03	7.4220E-05	1.9233E-02	2.0768E-04
		7	0.58426	3.2462E-04	1.8898E-06	4.6167E-03	1.1676E-05
		10	0.55523	1.8895E-05	7.7274E-09	5.7808E-04	1.8904E-07
		15	0.53637	1.8861E-07	8.0929E-13	2.0205E-05	2.3544E-10
		20	0.52940	1.9602E-09	8.4763E-17	7.9559E-07	3.0742E-13
		30	0.52445	2.1614E-13	9.2985E-25	1.7490E-09	5.3075E-19
		100	0.52145	4.4160E-41	1.7777E-80	1.1339E-26	2.4419E-59
		Min	**0.52144**	**0**	**0**	**0**	**0**

Conclusions and Future work

- **We have improved a set of similarity measures by using linguistic quantifier guided aggregation.**
- **These measures are also applicable when the variables are numeric (no uncertainty)**
- **The advantages of using RIM linguistic quantifiers to combine the individual similarities are:**
  - The aggregation is **soft** rather than **hard**, so we can tolerate some restrictions in the decision making
  - The measures can be easily adapted to the needs of each organization

- **The empirical validation of estimation effort by analogy must be achieved:**
  - For the individual distance, we use the two retained measures
  - For the overall distance, we use RIM linguistic quantifiers
- **Can I use our measures for prediction of Size, Reliability, Maintainability,…?**
- **Building prediction systems by analogy that satisfy Soft Computing:**
  - Tolerance of imprecision (Fuzzy Logic)
  - Learning (Neural Networks)
  - Uncertainty (Belief networks, genetic algorithms,…)