

# Vertex Page Clustering

Kaio Cesar Alves Reis, Matteo De Santis, Francesco Forgione

## Problema principale



- Sviluppo e implementazione dell'algoritmo di clustering di Vertex
- Dato un sito Web, l'algoritmo deve essere in grado di raggruppare le varie pagine del dominio in cluster
- Ciascun cluster fa riferimento ad un'entità
- L'algoritmo deve essere in grado di riconoscere l'entità descritta nella pagina Web



- L'idea di base di Vertex è quella di valutare le pagine Web in base alla loro struttura
- Per fare questo l'algoritmo viene suddiviso in tre passi



Nel primo passo viene implementata una hash table che contiene tutti i masked shingle vector, e il relativo punteggio, di tutte le pagine del dominio in esame

```
Input: Sample pages P from Web site to be clustered;
Output: Set of clusters;
/* First Pass */
Initialize hash table H to empty;
for each page p \in P do
  Let v be the shingle vector for p;
  for each 6/8, 7/8, and 8/8 masked shingle vector v'
  covering v do
     if v' is in hash table H then
       Increment the count for v';
     else
       Insert v' with count 1 into H;
     end if
  end for
end for
```



Nel secondo passo viene creato un set contenente tutti i masked shingle vector di ogni pagina con punteggio maggiore, i quali coprono gli shingle vector delle pagine campione

```
/* Second Pass */

for each 8/8 vector v in H in increasing order of counts do

Let v' be the masked shingle vector in H with maximum count covering v;

Decrement counts of all masked shingle vectors \neq v' in H covering v (by v's count);

end for

Delete masked shingle vectors with counts less than threshold from H;
```



Nel terzo passo, per ogni pagina del dominio vengono confrontati i masked shingle vector e successivamente questa viene assegnata ad un cluster

```
/* Third Pass */
For each masked shingle vector v \in H, C_v = \emptyset;
for each page p \in P with shingle vector v do

Let v' be the masked shingle vector in H with maximum count covering v;
Add p to C_{v'};
end for
return \{(C_v, v, count for v) : v \in H\};
```

### **Funzionamento**

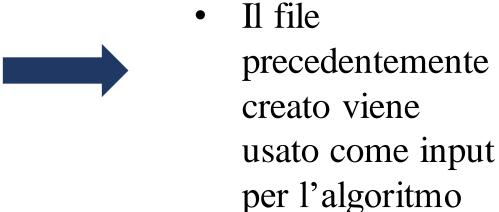


- Estrazione della ground truth relativa al dominio
- Per ciascuna pagina Web del dominio vengono estratti tutti i tag html
- I tag vengono utilizzati per creare tutti gli shingle di lunghezza 10
- Viene associata una firma ad ogni pagina Web creata in base a tutti gli shingle di questa
- Viene creato un file contenente pagine e relative firme

### Funzionamento



### Input



di clustering

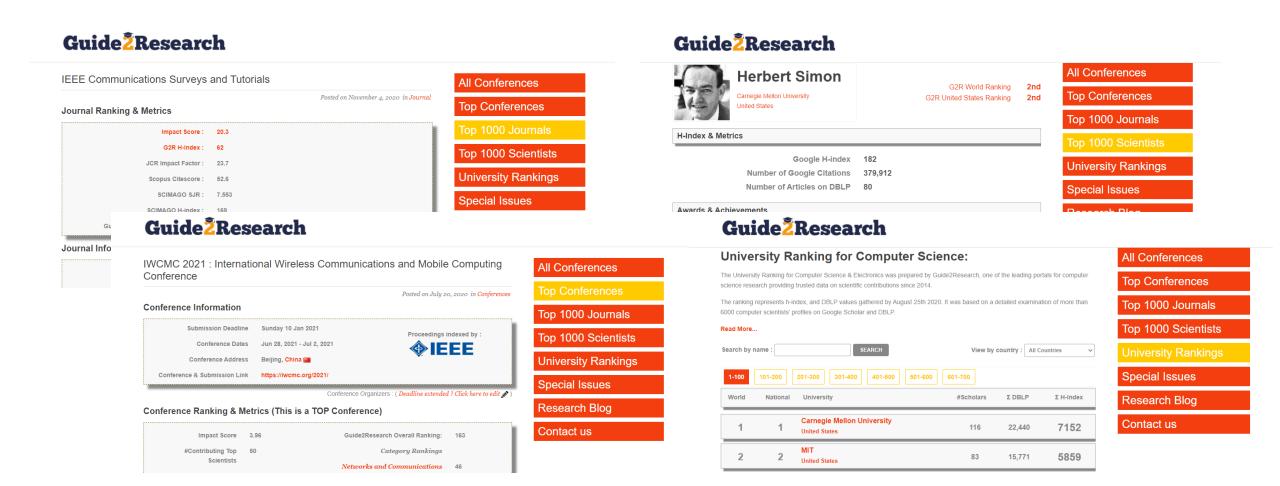
### Output

 L'output del programma è un file contenente tutti i cluster individuati



### Metriche: Guide2Research





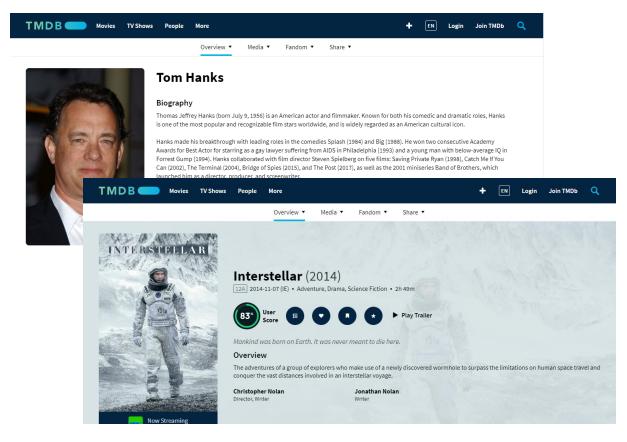
## Metriche: Guide2Research

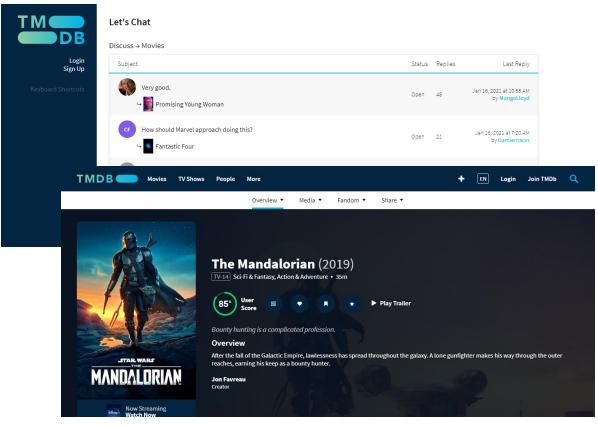


Cluster	Precision	Recall	<b>F1</b>
Conference	0,96	1	0,98
Journal	1	1	1
Special Issue	0,79	1	0,88
User	1	1	1

### Metriche: The Movie DB







### Metriche: The Movie DB



Cluster	Precision	Recall	F1
Discover	1	0,74	0,84
Movie	0,78	0,93	0,90
Network	0,94	1	0,98
Person	0,96	1	0,97
Review	0,76	1	0,88
Talk	0,43	0,9	0,71
Tv	0,82	0,92	0,78
U	0,96	0,99	0,96

### Metriche: Study.eu



### Study Innovation Management: All you need to know

Author: Megan Thudium - Last update: 27 August 2020

Innovation Management is a versatile and exciting degree to study at university. It's the perfect subject if you are forward-thinking and creative and enjoy combining high-level strategy with hands-on work on new business ideas. Students that pursue this field are well-equipped to become the innovators, the thinkers, the creatives, and the leaders of tomorrow. Read on to learn more:

### What is Innovation Management?

Innovation Management is a field promoting organizational innovation and includes processes that encourage an organization to strategically innovate. Studying Innovation Management at university will equip you to solve the common challenges organizations face to successfully innovate, whether that company is a corporation, enterprise, start-up, or small business.

A university degree in Innovation Management will provide you with insight into the role that innovation plays in organizations, how it ties into corporate strategy, product development, market positioning, and new

Innovation Managers perform tasks like

- Promote innovation and innovative processes throughout an orga
- Develop and collect ideas for innovation.
- Work directly with leadership to make strategic long-term plans.

As a graduate of Innovation Management, you can fundamentally trans and company cultures. Ready to make an impact? If yes, then a university



Berlin, Germany

4 study programmes



### **University of Kent**

The University is one of the country's leading academic institutions producing worldclass research, rated internationally excellent and leading the way in many fields of

We are a forward-thinking research institution, committed to the transformative power of education and research and to the development and support of our students and staff.



### Bachelors at the University of Kent

### University of **University of Kent** Canterbury (United Location: Kingdom) 1965 Founded: 19 300 (17% international Students: Academic staff: 2,000 #170 (Europe) Ranking: 6 #351 (World) Listed study programmes Website: kent.ac.uk ☑



### Italy

Italy's historic treasures and magnificent cuisine continue to draw visitors as they have done for centuries. From the snow-capped Dolomites in the North to the warm maritime Mediterranean culture of the South, Italy's regions are diverse and exciting. For the international student seeking world-class but affordable education, with a wide range of social and leisure activities, few countries are as exciting a choice as Italy.

The country's cultural importance extends back to ancient times and the past greets the student of history and archaeology on every street corner. For many, Rome is still the Eternal City, the centre of an empire that - some 2,000 years ago - stretched from rainy Britannia to the deserts of Syria. At its height it also encompassed Egypt, much of North Africa and nearly all of Continental Europe.

Italy's Ancient Roman monuments, magnificent as they are, tell only part of the story. The Renaissance of the 14th to 17th century originated in Italy (as the Rinascimento) and sparked innovation and revolutions in art, architecture, philosophy, religion and the sciences across Europe. Intellectual activity and creativity flourished, carrying gems of ancient and medieval thought into the modern world.

### Studying at a university in Italy

It's perhaps not surprising that many international students feel instantly at home in Italy. The University of Bologna has been welcoming students since 1088 – it is the oldest university in the world and its foundation gave the city one of its nicknames: la dotta, "the learned one". Bologna remains at the leading edge of academic innovation in the 21st century. The Bologna Process was initiated in the city, and has established a set of standards for excellence in higher education across the European Higher Education Area (EHEA).

Italian universities take pride in their heritage and rankings. The University of Rome Tor Vergata offers more than a dozen international degree courses that are taught in English. Founded only in 1982, Tor Vergata was ranked one of



### FEATURED UNIVERSITY:





## Metriche: Study.eu



Cluster	Precision	Recall	F1
Faculty	0,92	1	0,95
Search	0,87	1	0,93
University	0,54	1	0,70

### Considerazioni



- Il raggruppamento delle pagine nel dataset è stato fatto manualmente
- L'algoritmo risulta avere una complessità n²
- Processa e classifica circa 15 pagine al secondo
- L'esperimento è stato eseguito con una threshold pari a 10
- Con valori di threshold maggiori il numero di cluster individuati è minore