Assegnamento Automatico Di Macrocategorie Agli Articoli Di Wikipedia

Tesi di Laurea Triennale

Tesi di: Jacopo Farina, matricola 713091

Relatore: Prof. Marco Colombetti
Correlatori: Ing. Riccardo Tasso, Ing. David Laniado

Sessione di laurea del 21 settembre 2010

Wikipedia

- Enciclopedia consultabile via Web il cui contenuto è **liberamente modificabile** da chiunque
- Lanciata il 15 gennaio 2001, ha avuto una crescita esplosiva
- Disponibile in oltre 270 lingue, di cui un centinaio attive
- La versione in inglese conta più di **3.4 milioni** di voci, ed è la più grande in assoluto seguita da quella in tedesco con 1.2 milioni
- La versione in italiano ne ha "solo" 720.000

Le categorie di Wikipedia

Nel maggio 2004 sono state introdotte le categorie per organizzare meglio la quantità sempre maggiore di articoli esistenti.

- Attualmente ne esistono circa 500mila
- Ogni pagina può appartenere a una o più categorie, scelte arbitrariamente dagli utenti
  - In media una pagina appartiene a **2,68** categorie
  - Il 64% degli articoli appertiene a una o due categorie, il 93% a meno di 7
  - La pagina *Winston Churchill* ha il numero record di 70 categorie di appartenenza, *Albert Einstein* ne ha 56.
- Ogni categoria appartiene ad altre categorie e ne può contenere altre, senza vincoli particolari
- **non rappresentano una tassonomia**, anche se di solito sono organizzate in maniera gerarchica, semmai dei legami semantici

Le categorie di Wikipedia

Nel maggio 2004 sono state introdotte le categorie per organizzare meglio la quantità sempre maggiore di articoli esistenti.

- Attualmente ne esistono circa 500mila
- Ogni pagina può appartenere a una o più categorie, scelte arbitrariamente dagli utenti
  - In media una pagina appartiene a **2,68** categorie
  - Il 64% degli articoli appertiene a una o due categorie, il 93% a meno di 7
  - La pagina *Winston Churchill* ha il numero record di 70 categorie di appartenenza, *Albert Einstein* ne ha 56.
- Ogni categoria appartiene ad altre categorie e ne può contenere altre, senza vincoli particolari
- **non rappresentano una tassonomia**, anche se di solito sono organizzate in maniera gerarchica, semmai dei legami semantici

Le categorie di Wikipedia

Nel maggio 2004 sono state introdotte le categorie per organizzare meglio la quantità sempre maggiore di articoli esistenti.

- Attualmente ne esistono circa 500mila
- Ogni pagina può appartenere a una o più categorie, scelte arbitrariamente dagli utenti
  - In media una pagina appartiene a **2,68** categorie
  - Il 64% degli articoli appertiene a una o due categorie, il 93% a meno di 7
  - La pagina *Winston Churchill* ha il numero record di 70 categorie di appartenenza, *Albert Einstein* ne ha 56.
- Ogni categoria appartiene ad altre categorie e ne può contenere altre, senza vincoli particolari
- **non rappresentano una tassonomia**, anche se di solito sono organizzate in maniera gerarchica, semmai dei legami semantici

Una pagina di Wikipedia

Milan

From Wikipedia, the free encyclopedia

Coordinates: 45°27′51″N 09°

"Milano" redirects here. For other uses, see Milano (disambiguation).

For other uses, see Milan (disambiguation).

Milan (Italian: Milano, About this sound listen (help·info) Italian pronunciation: [mi'la(:)no]; Western Lombard: Milan, About this sound listen (help·info)) is a city in Italy and the capital of the region of Lombardy and of the province of Milan. The city proper has a population of about 1,310,000, while the urban area is the first in Italy and the fifth largest in the European Union with a population of 4,345,000 over an area of 2,370 km2 (915 sq mi).[2] The Milan metropolitan area, by far the largest in Italy, is estimated by the OECD to have a population of 7,400,000.[3]

- The Milan Garden of the Righteous
- Milan in Hebrew מילאנו

v · d · e **Regional capitals of Italy**

v · d · e **Lombardy - *Comuni* of the Province of Milan**

Categories:Cities and town in Lombardy|Communes of the province of Milan|Populated places estabilished in 1st millennium

Una categoria di Wikipedia

Category Discussion Read Edit View history Search

Category:Populated places established in the 1st millennium BC

From Wikipedia, the free encyclopedia

Wikimedia Commons has media related to: *Settlements established in the 1st millennium BC*

Subcategories

This category has the following 10 subcategories, out of 10 total.

- [×] Populated places established in the 1st century BC (13 P)
- [+] Populated places established in the 2nd century BC (2 C, 13 P)
- [×] Populated places established in the 3rd century BC (18 P)
- [×] Populated places established in the 4th century BC (33 P)

cont.

- [×] Populated places established in the 5th century BC (14 P)
- [×] Populated places established in the 6th century BC (16 P)
- [×] Populated places established in the 7th century BC (22 P)
- [×] Populated places established in the 8th century BC (16 P)

cont.

- [×] Populated places established in the 9th century BC (6 P)
- [×] Populated places established in the 10th century BC (1 P)

Pages in category "Populated places established in the 1st millennium BC"

The following 28 pages are in this category, out of 28 total. This list may not reflect recent changes (learn more).

G

- Geneva

H

- Herculaneum

K

- Ksar-el-Kebir

M

- Marseille

S

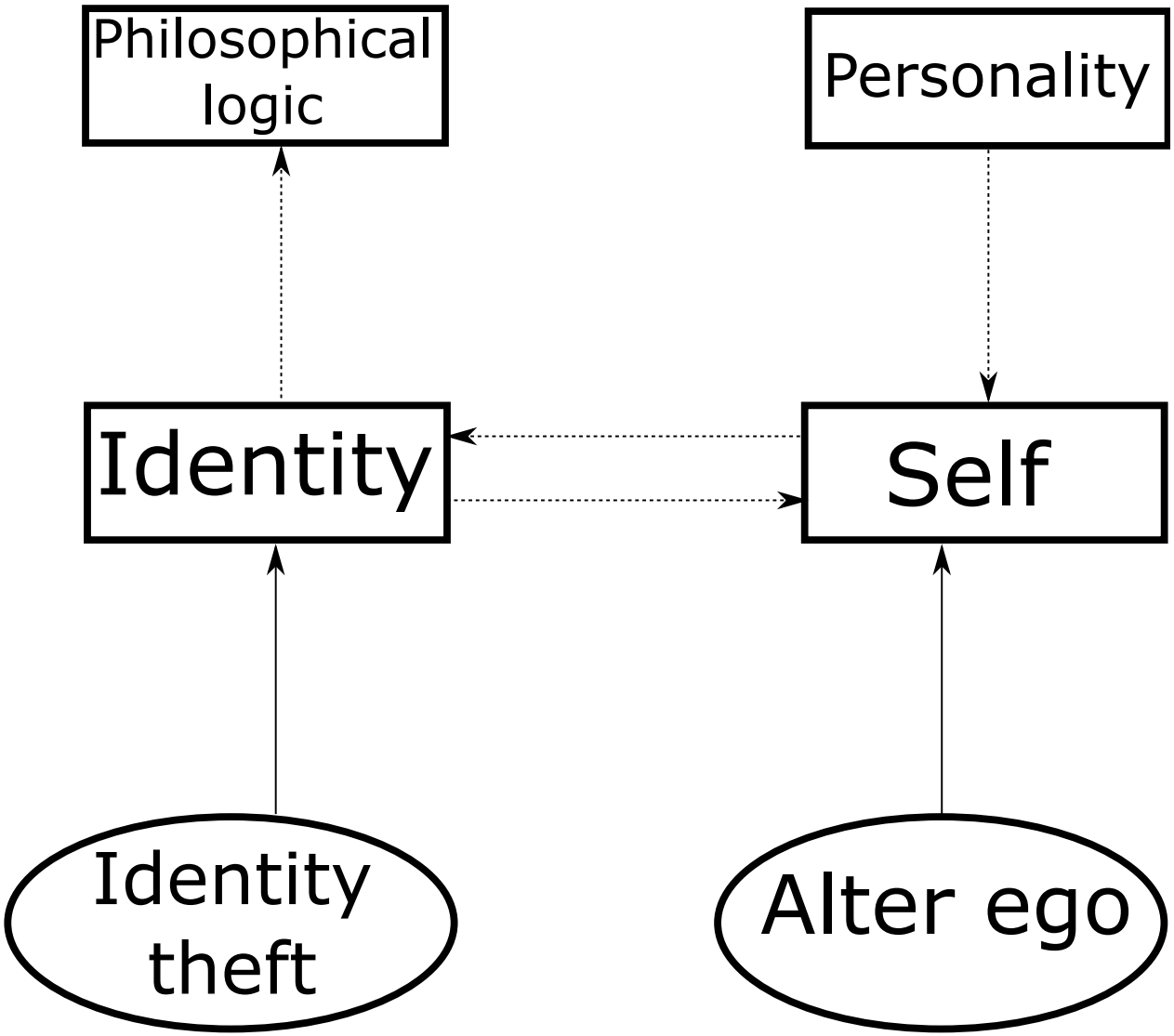
- Seleucia Pieria
- Siena
- Soli, Cilicia

Y

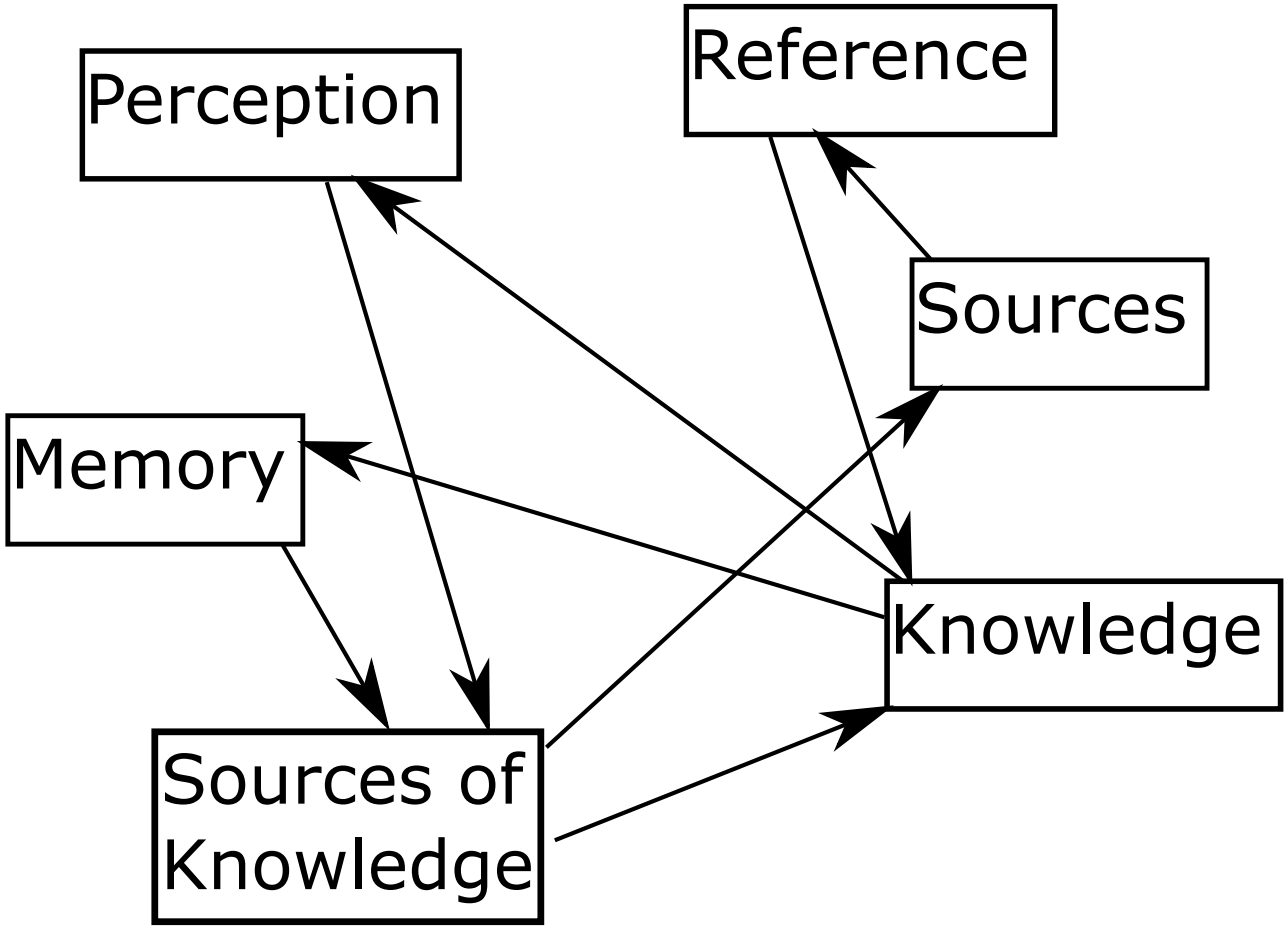
- Yangzhou
- Yevpatoria

Categories: Populated places by year of establishment | 1st-millennium BC establishments

Un esempio reale



Un ciclo interessante



- Gli archi non rappresentano sempre una relazione di tipo ***isa***
- I cicli si trovano con l'algoritmo di Tarjan per le strutture fortemente connesse

Utilizzi nell'IA

Wikipedia contiene quantità **enormi** di testo **liberamente accessibile** su qualsiasi argomento, ed è quindi di grande interesse nel campo dell'Intelligenza Artificiale:

- Mappatura tra termini di ontologie (come **Cyc**) e articoli di Wikipedia per estrarre conoscenza
- Analisi automatica del testo per estrarre legami tra parole (sinonimi, iperonimi, iponimi...)
- Ricerca di legami semantici tra i termini
- Analisi statistica degli argomenti trattati dal progetto e dai singoli utenti

Utilizzi nell'IA

Wikipedia contiene quantità **enormi** di testo **liberamente accessibile** su qualsiasi argomento, ed è quindi di grande interesse nel campo dell'Intelligenza Artificiale:

- Mappatura tra termini di ontologie (come **Cyc**) e articoli di Wikipedia per estrarre conoscenza
- Analisi automatica del testo per estrarre legami tra parole (sinonimi, iperonimi, iponimi...)
- Ricerca di legami semantici tra i termini
- Analisi statistica degli argomenti trattati dal progetto e dai singoli utenti

Utilizzi nell'IA

Wikipedia contiene quantità **enormi** di testo **liberamente accessibile** su qualsiasi argomento, ed è quindi di grande interesse nel campo dell'Intelligenza Artificiale:

- Mappatura tra termini di ontologie (come **Cyc**) e articoli di Wikipedia per estrarre conoscenza
- Analisi automatica del testo per estrarre legami tra parole (sinonimi, iperonimi, iponimi...)
- Ricerca di legami semantici tra i termini
- Analisi statistica degli argomenti trattati dal progetto e dai singoli utenti

Utilizzi nell'IA

Wikipedia contiene quantità **enormi** di testo **liberamente accessibile** su qualsiasi argomento, ed è quindi di grande interesse nel campo dell'Intelligenza Artificiale:

- Mappatura tra termini di ontologie (come **Cyc**) e articoli di Wikipedia per estrarre conoscenza
- Analisi automatica del testo per estrarre legami tra parole (sinonimi, iperonimi, iponimi...)
- Ricerca di legami semantici tra i termini
- Analisi statistica degli argomenti trattati dal progetto e dai singoli utenti

Utilizzi nell'IA

Wikipedia contiene quantità **enormi** di testo **liberamente accessibile** su qualsiasi argomento, ed è quindi di grande interesse nel campo dell'Intelligenza Artificiale:

- Mappatura tra termini di ontologie (come **Cyc**) e articoli di Wikipedia per estrarre conoscenza
- Analisi automatica del testo per estrarre legami tra parole (sinonimi, iperonimi, iponimi...)
- Ricerca di legami semantici tra i termini
- Analisi statistica degli argomenti trattati dal progetto e dai singoli utenti

Il problema dell'assegnamento

- Dato un set di argomenti, chiamati **macrocategorie**, che partiziona idealmente la conoscenza umana, come stabilire automaticamente quale (o quali) di essi è più adatto a contenere un certo articolo ?
- Kittur (2008) ha analizzato il grafo delle categorie di Wikipedia come se fosse una **rete semantica**, abbinando ogni articolo alla macrocategoria con la minore distanza topologica dall'articolo tra quelle di un set di 11 elementi.
- Il criterio di Kittur è ancora valido dopo l'aumento delle dimensioni del grafo? E se si scelgono più macrocategorie?
- Esistono criteri più efficaci, ossia che assegnino automaticamente le macrocategorie agli articoli in maniera più simile a quanto farebbe un essere umano?

Il problema dell'assegnamento

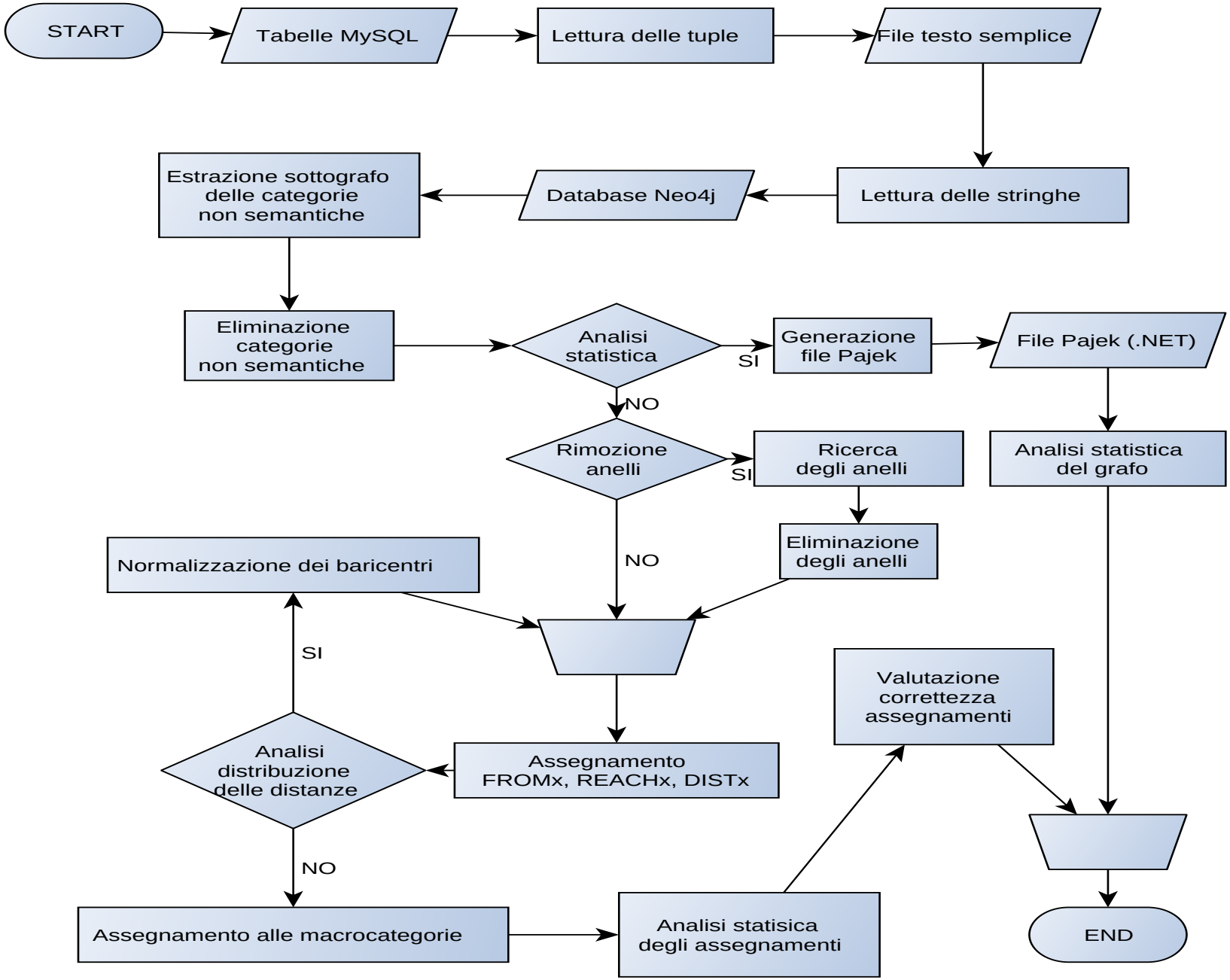
- Dato un set di argomenti, chiamati **macrocategorie**, che partiziona idealmente la conoscenza umana, come stabilire automaticamente quale (o quali) di essi è più adatto a contenere un certo articolo ?
- Kittur (2008) ha analizzato il grafo delle categorie di Wikipedia come se fosse una **rete semantica**, abbinando ogni articolo alla macrocategoria con la minore distanza topologica dall'articolo tra quelle di un set di 11 elementi.
- Il criterio di Kittur è ancora valido dopo l'aumento delle dimensioni del grafo? E se si scelgono più macrocategorie?
- Esistono criteri più efficaci, ossia che assegnino automaticamente le macrocategorie agli articoli in maniera più simile a quanto farebbe un essere umano?

Il problema dell'assegnamento

- Dato un set di argomenti, chiamati **macrocategorie**, che partiziona idealmente la conoscenza umana, come stabilire automaticamente quale (o quali) di essi è più adatto a contenere un certo articolo ?
- Kittur (2008) ha analizzato il grafo delle categorie di Wikipedia come se fosse una **rete semantica**, abbinando ogni articolo alla macrocategoria con la minore distanza topologica dall'articolo tra quelle di un set di 11 elementi.
- Il criterio di Kittur è ancora valido dopo l'aumento delle dimensioni del grafo? E se si scelgono più macrocategorie?
- Esistono criteri più efficaci, ossia che assegnino automaticamente le macrocategorie agli articoli in maniera più simile a quanto farebbe un essere umano?

Il problema dell'assegnamento

- Dato un set di argomenti, chiamati **macrocategorie**, che partiziona idealmente la conoscenza umana, come stabilire automaticamente quale (o quali) di essi è più adatto a contenere un certo articolo ?
- Kittur (2008) ha analizzato il grafo delle categorie di Wikipedia come se fosse una **rete semantica**, abbinando ogni articolo alla macrocategoria con la minore distanza topologica dall'articolo tra quelle di un set di 11 elementi.
- Il criterio di Kittur è ancora valido dopo l'aumento delle dimensioni del grafo? E se si scelgono più macrocategorie?
- Esistono criteri più efficaci, ossia che assegnino automaticamente le macrocategorie agli articoli in maniera più simile a quanto farebbe un essere umano?


START
Tabelle MySQL
Lettura delle tuple
File testo semplice
Lettura delle stringhe
Database Neo4j
Estrazione sottografo delle categorie non semantiche
Eliminazione categorie non semantiche
Analisi statistica
SI
Generazione file Pajek
File Pajek (.NET)
NO
Rimozione anelli
SI
Ricerca degli anelli
Analisi statistica del grafo
Eliminazione degli anelli
NO
Normalizzazione dei baricentri
SI
Valutazione correttezza assegnamenti
Analisi distribuzione delle distanze
Assegnamento FROMx, REACHx, DISTx
NO
Assegnamento alle macrocategorie
Analisi statisica degli assegnamenti
END

Formato dei risultati e valutazione

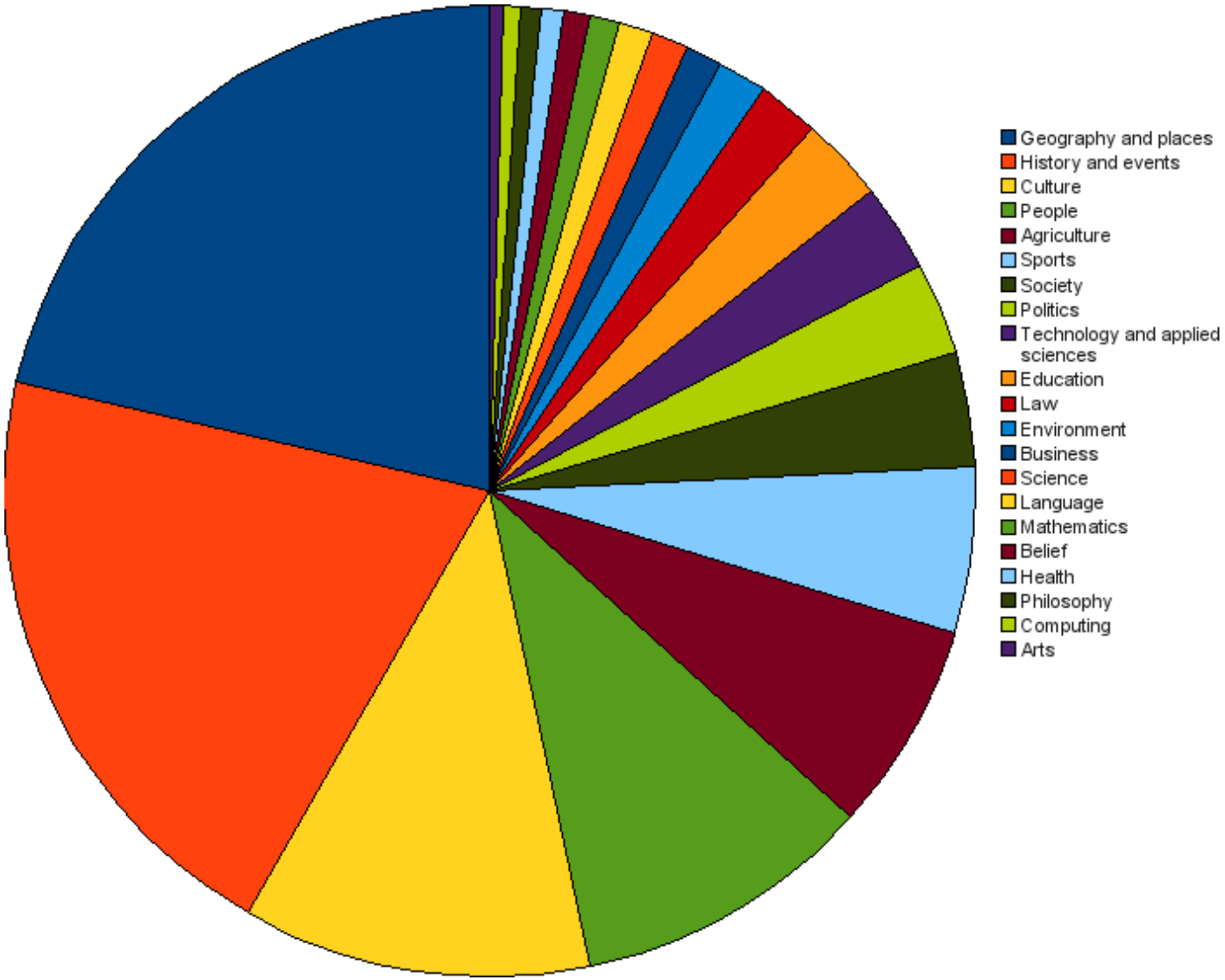
- Per ogni pagina ottengo delle percentuali di assegnamento alle macrocategorie

  Politecnico_di_Milano: Education:87,5; History and events:12,5;

- Valuto la precisione dei risultati confrontando le quote assegnate con quelle stabilite manualmente da un umano
- Valutazione di 50 pagine scelte a caso tra quelle assegnate
- Viene usato come valore di similarità il coseno dei vettori degli assegnamenti (automatico ed umano)

$$cos(A, B) = \frac{A \bullet B}{||A|| * ||B||}$$

Dimensione delle macrocategorie

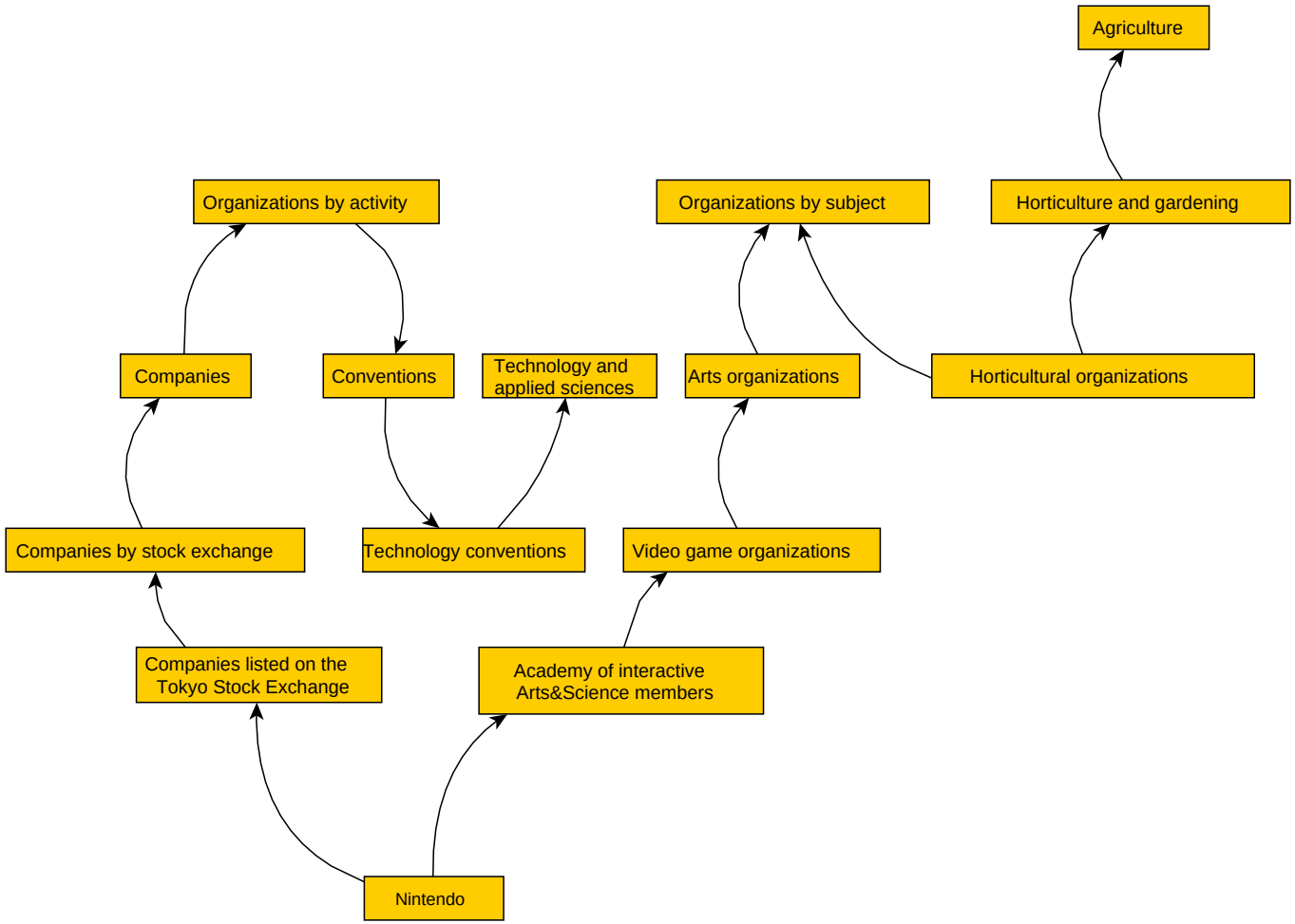


Alcuni assegnamenti

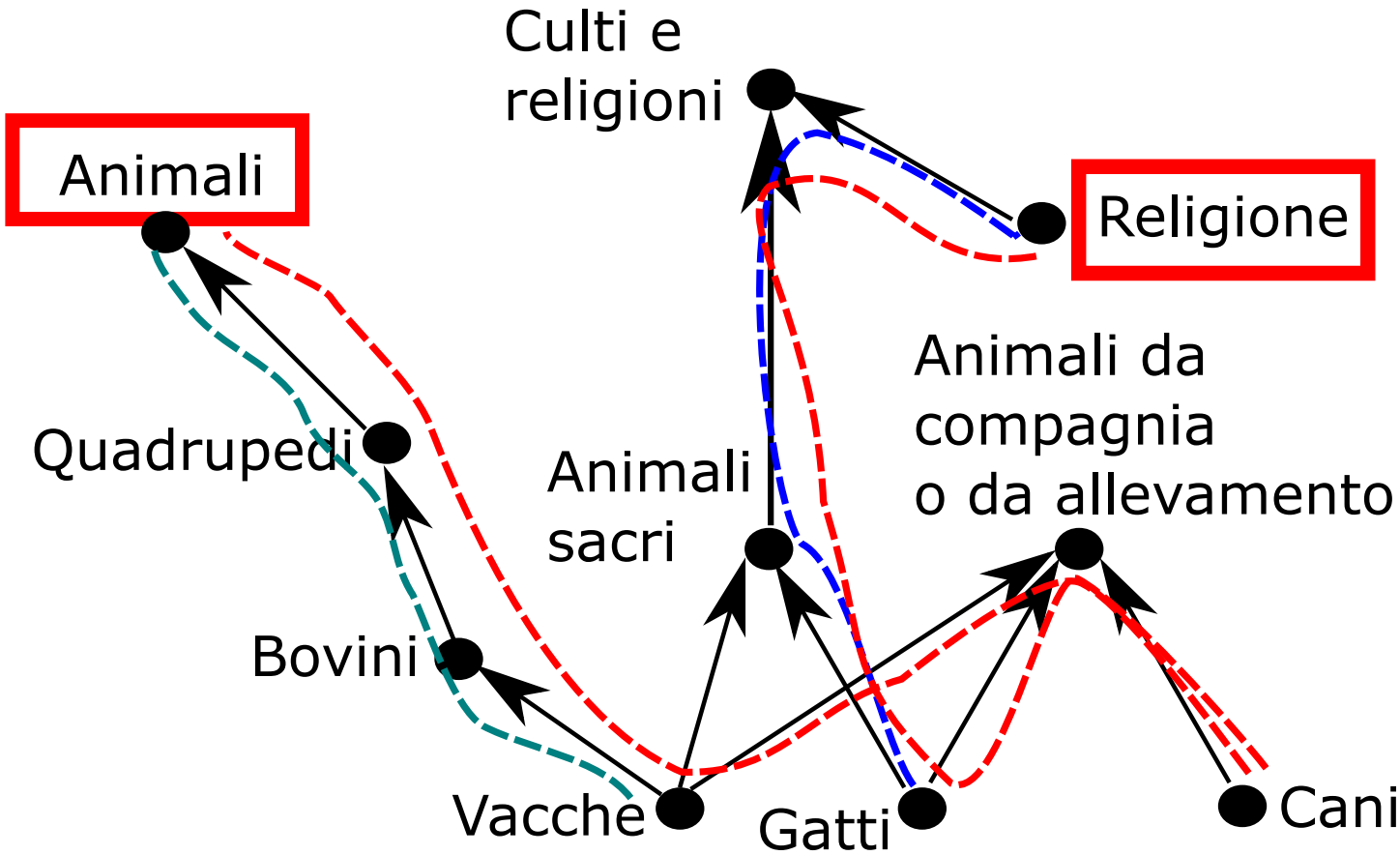
Milan: History and events:100;
Italy: History and events:70; Culture:30;
Politecnico_di_Milano: Education:87,5; History and events:12,5;
Java(software_platform): Computing:50; Technology and applied sciences:50;
Java: Geography and places:100; (L'isola principale dell'Indonesia)
Earth: Geography and places:100;
Albert_Einstein: Technology and applied sciences:18,53; Science:18,53; People:24,78; Mathematics:18,53; Politics:11,39; History and events:5,31; Society:2,93;
Nintendo: History and events:71,67; Agriculture:6,67; Culture:6,67; Geography and places:6,67; Philosophy:6,67; Technology and applied sciences:1,67; (famosa azienda produttrice di videogiochi)

- Perché Nintendo viene assegnata ad Agriculture più che a Technology and applied sciences ?

I percorsi da Nintendo

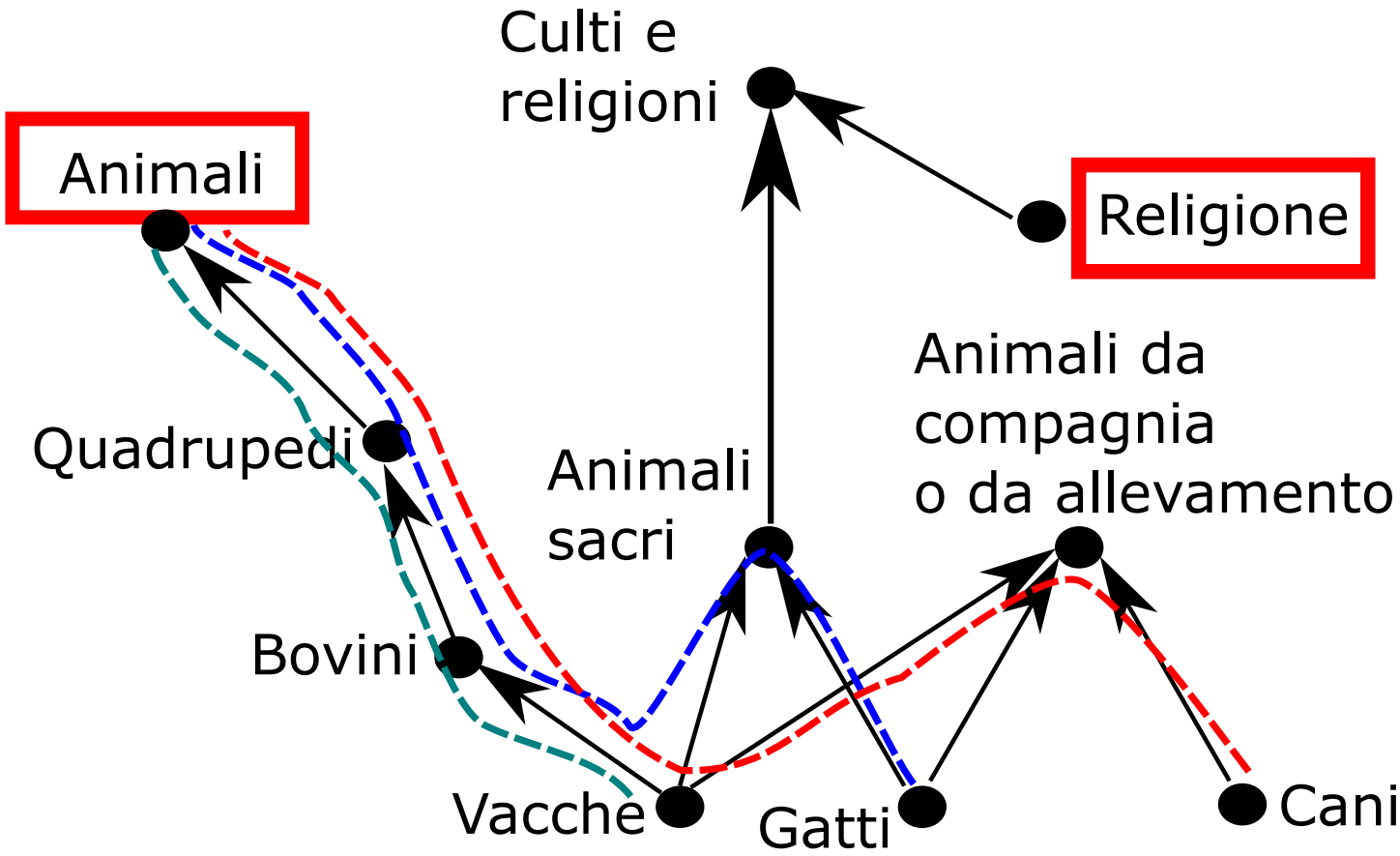


Percorsi diretti e indiretti



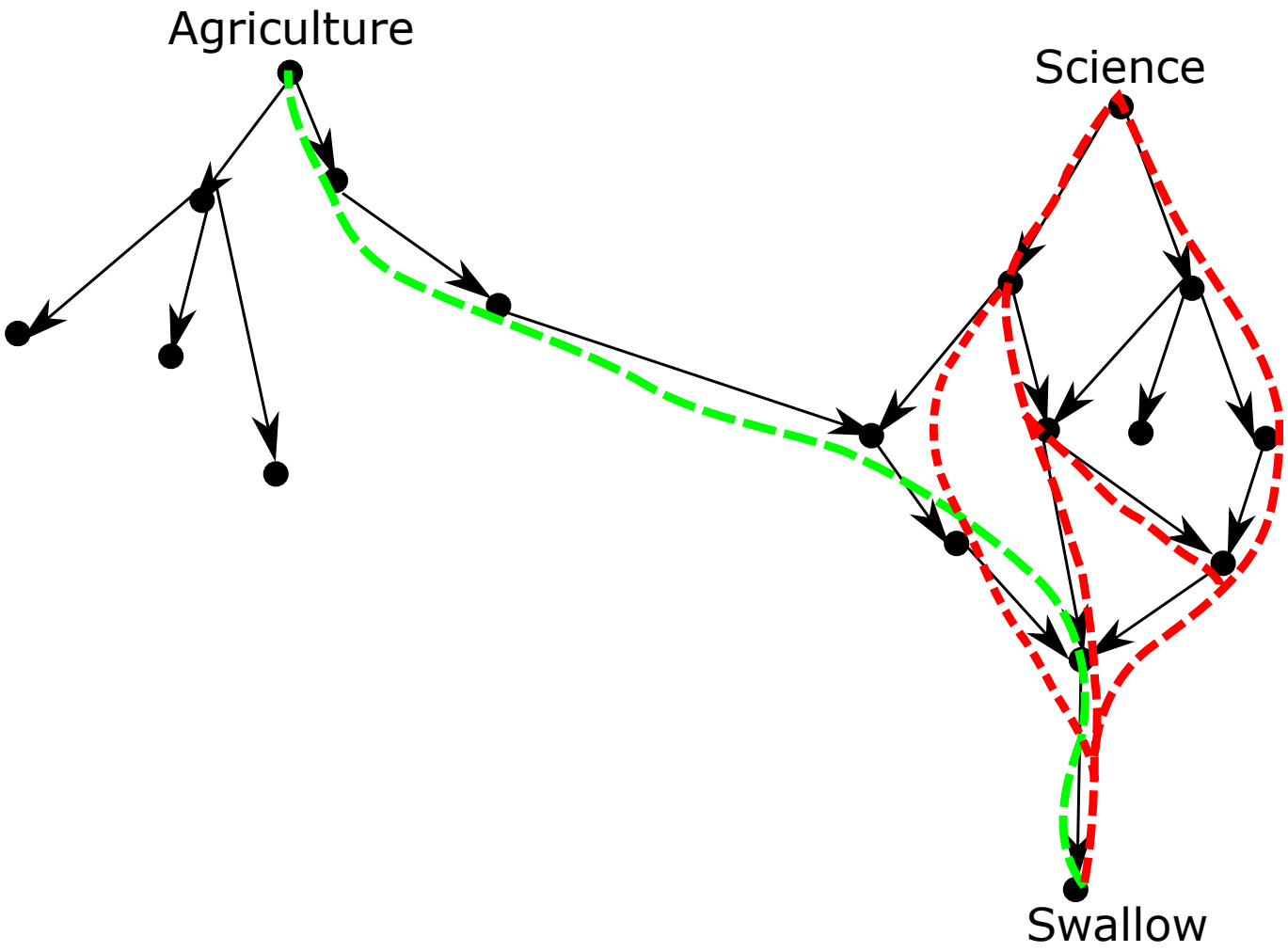
- Percorsi orientati: non sempre arrivo a una macrocategoria
- Percorsi non orientati: posso giungere a degli errori

Distribuzione con i costi differenziati



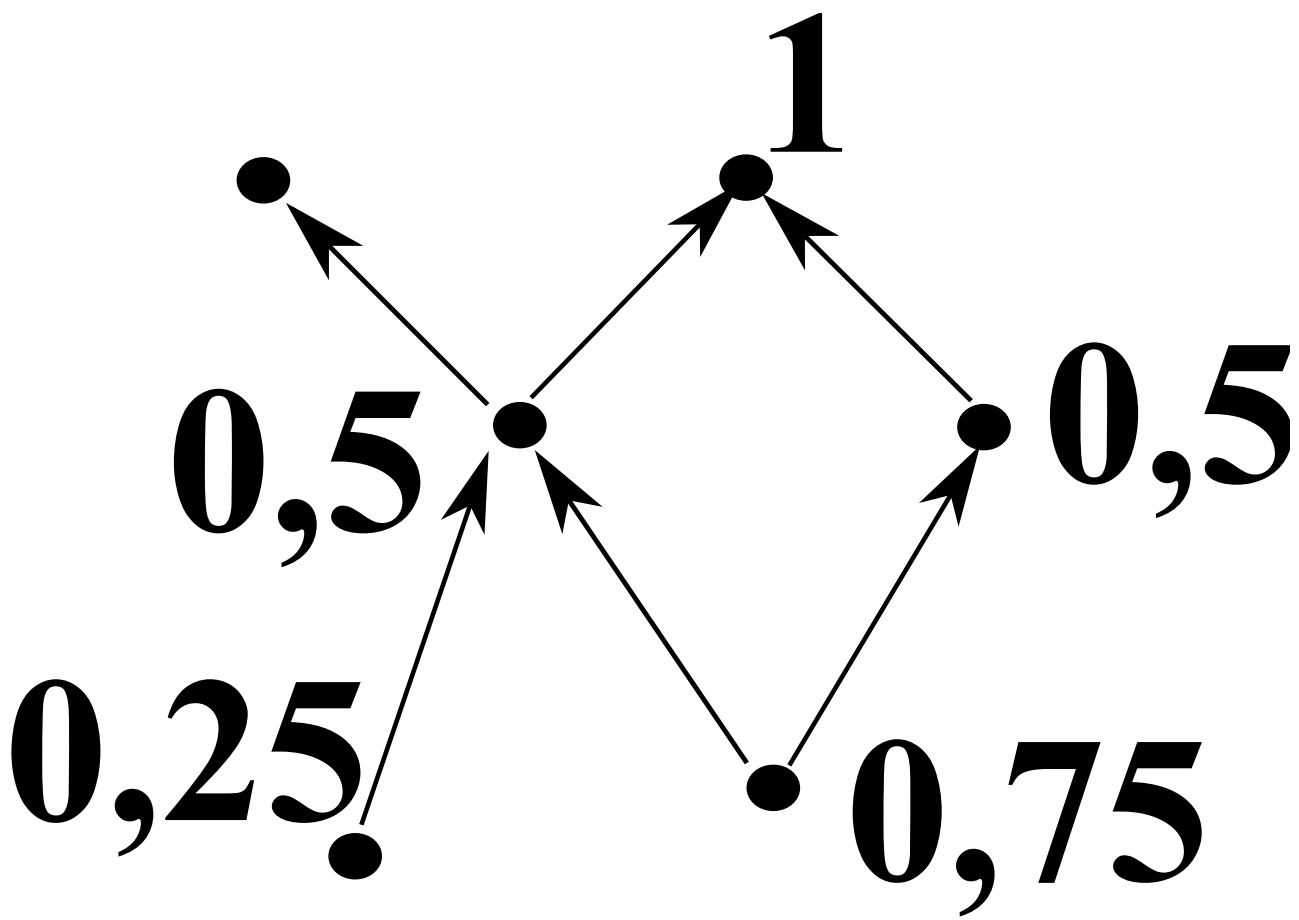
- Ibrido: mi muovo in entrambi i modi preferendo i percorsi orientati

Considerare la connettività



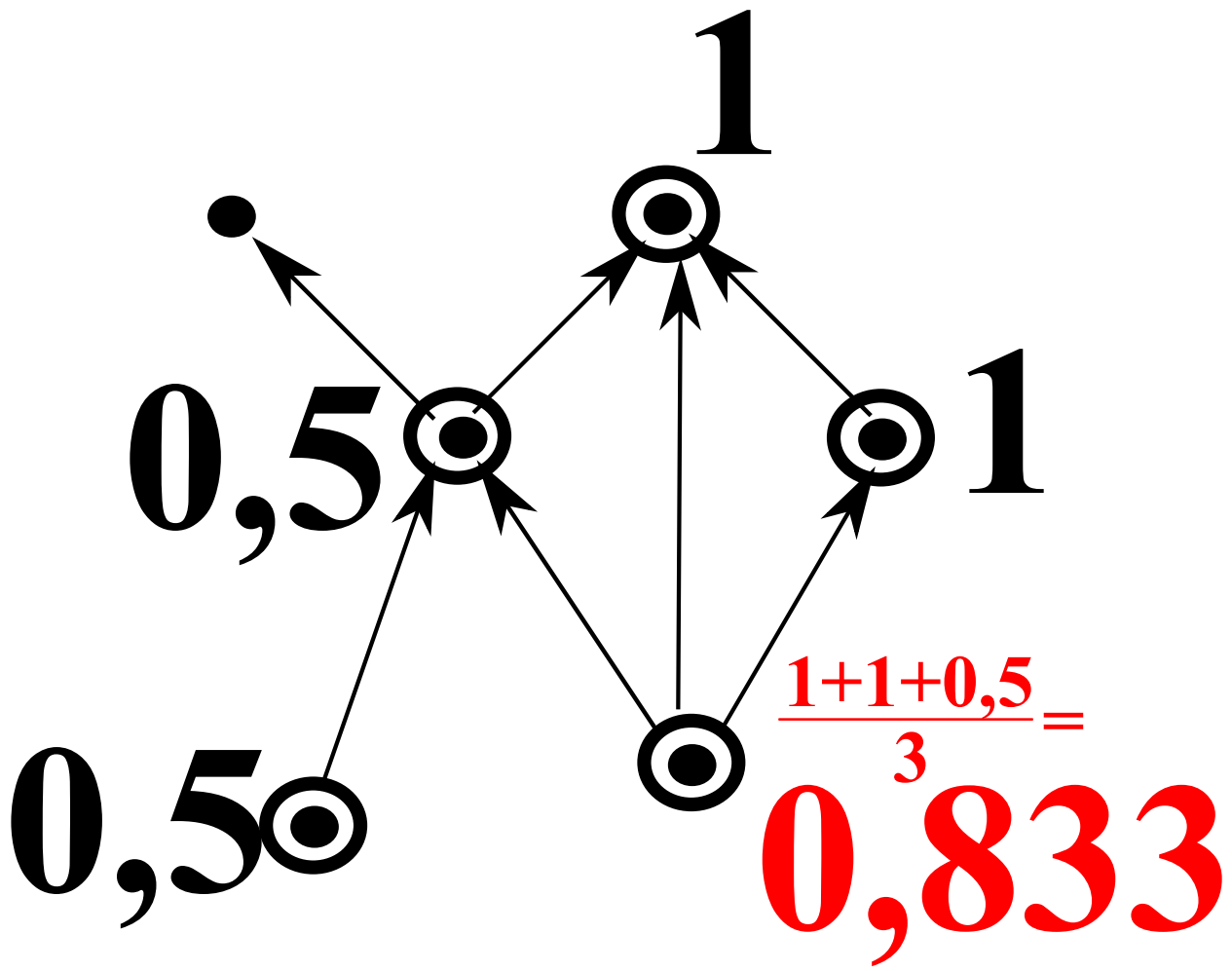
Potrebbe essere utile considerare più percorsi possibili alla volta.

Ripartizione di punteggi



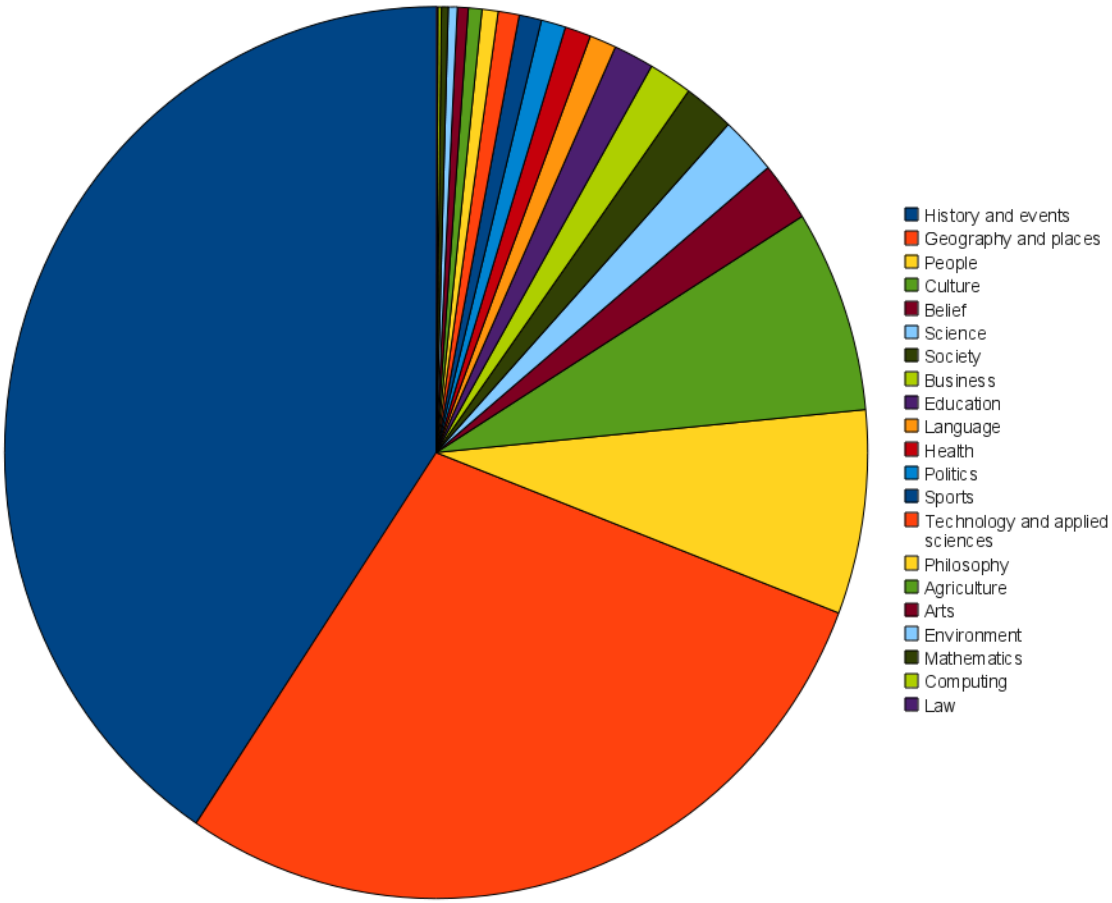
- Ripartisco il punteggio di una categoria tra quelle contenute, eventualmente sommandolo a quelli già assegnati.

Probabilità di raggiungimento



Marco i nodi raggiungibili, ogni categoria ha una probabilità uguale alla somma delle probabilità di quelle che la contengono divisa per gli archi uscenti.

Dimensione delle macrocategorie con assegnamento probabilistico



- Si nota che Agriculture si è molto ridimensionata, assumendo una quota più plausibile.

I metodi esaminati

Metodo	valore di correttezza	% pagine valutate
Caso base (Kittur, Holloway)	0.34	100%
Solo percorsi diretti	0.35	65%
Costi differenziati	**0.37**	100%
Ripartizione dei punteggi	0.35	62%
Probabilità di raggiungimento	0.36	62%

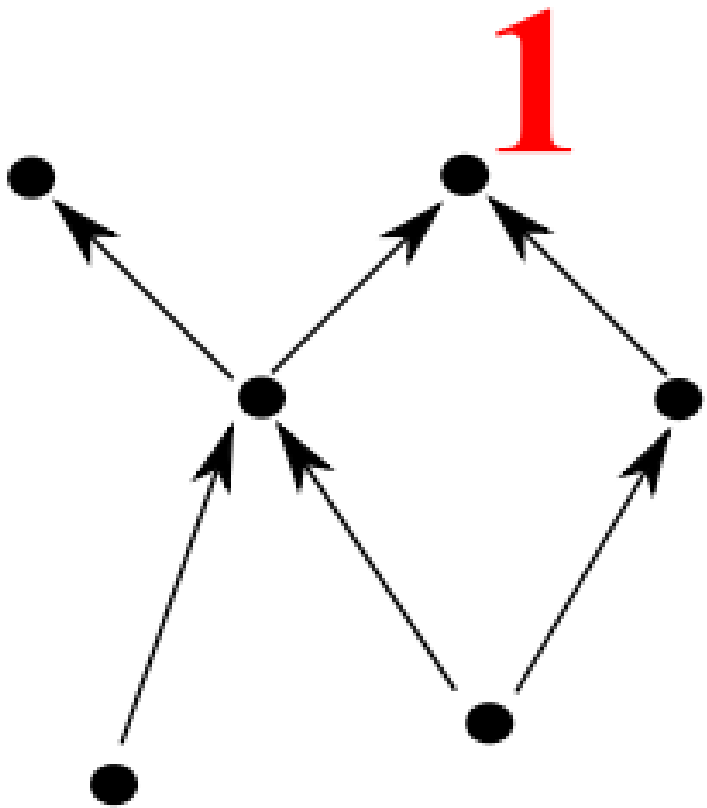
Conclusioni

- Il metodo più efficace tra quelli esaminati è quello a costi di attraversamento differenziati.
- Anche gli algoritmi basati sui percorsi multipli funzionano meglio del criterio basato sulla semplice distanza topologica
- Il metodo di Kittur fornisce assegnamenti correlati a quelli forniti manualmente da un valutatore umano anche se aumenta la complessità del grafo e si scelgono più macrocategorie.

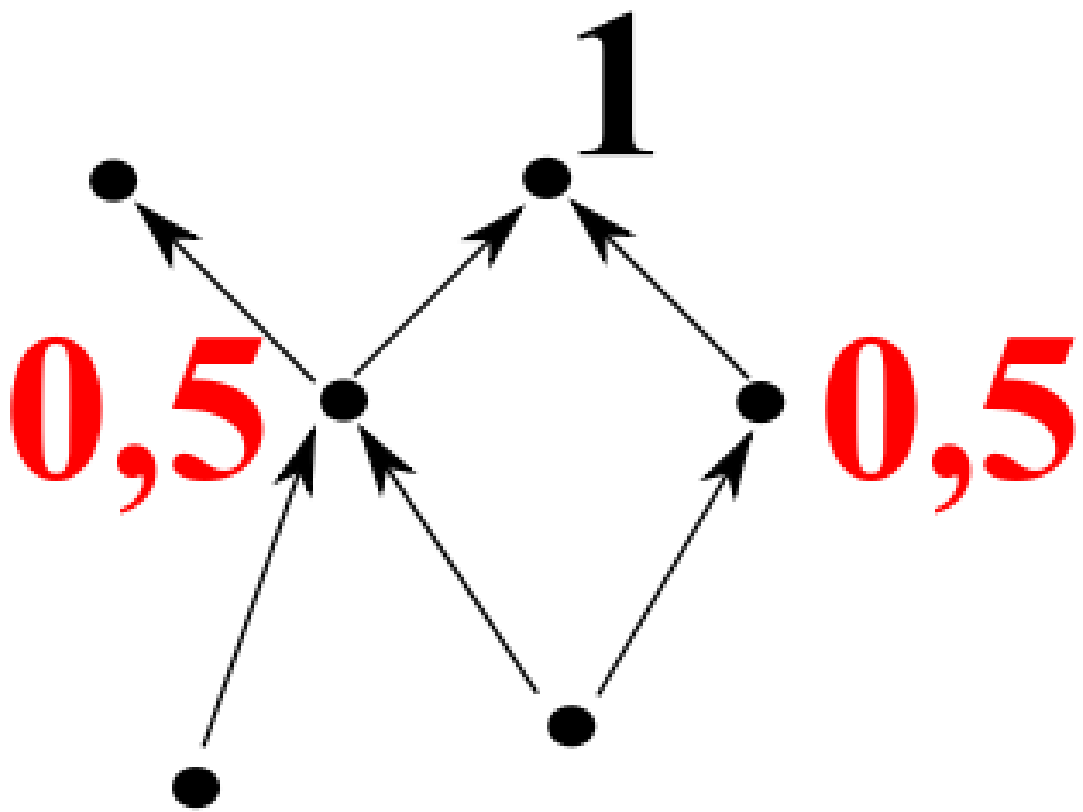
Sviluppi futuri

- Utilizzare non solo le appartenenze alle categorie ma anche i collegamenti tra le pagine per generare il grafo
- Elaborare tecniche del calcolo dei percorsi multipli che classifichino il 100% delle pagine seguendo gli archi in tutte le direzioni
  - Molte delle tecniche del calcolo del flusso esistenti non sono applicabili su grafi così grossi
  - Si potrebbero utilizzare approcci stocastici per calcolare approssimativamente la raggiungiblità di un nodo
- Utilizzare gli assegnamenti ottenuti per applicazioni pratiche
  - Mappare gli articoli con documenti testuali per calcolarne la vicinanza semantica
  - Stabilire la macrocategoria di un documento analizzandone le singole parole

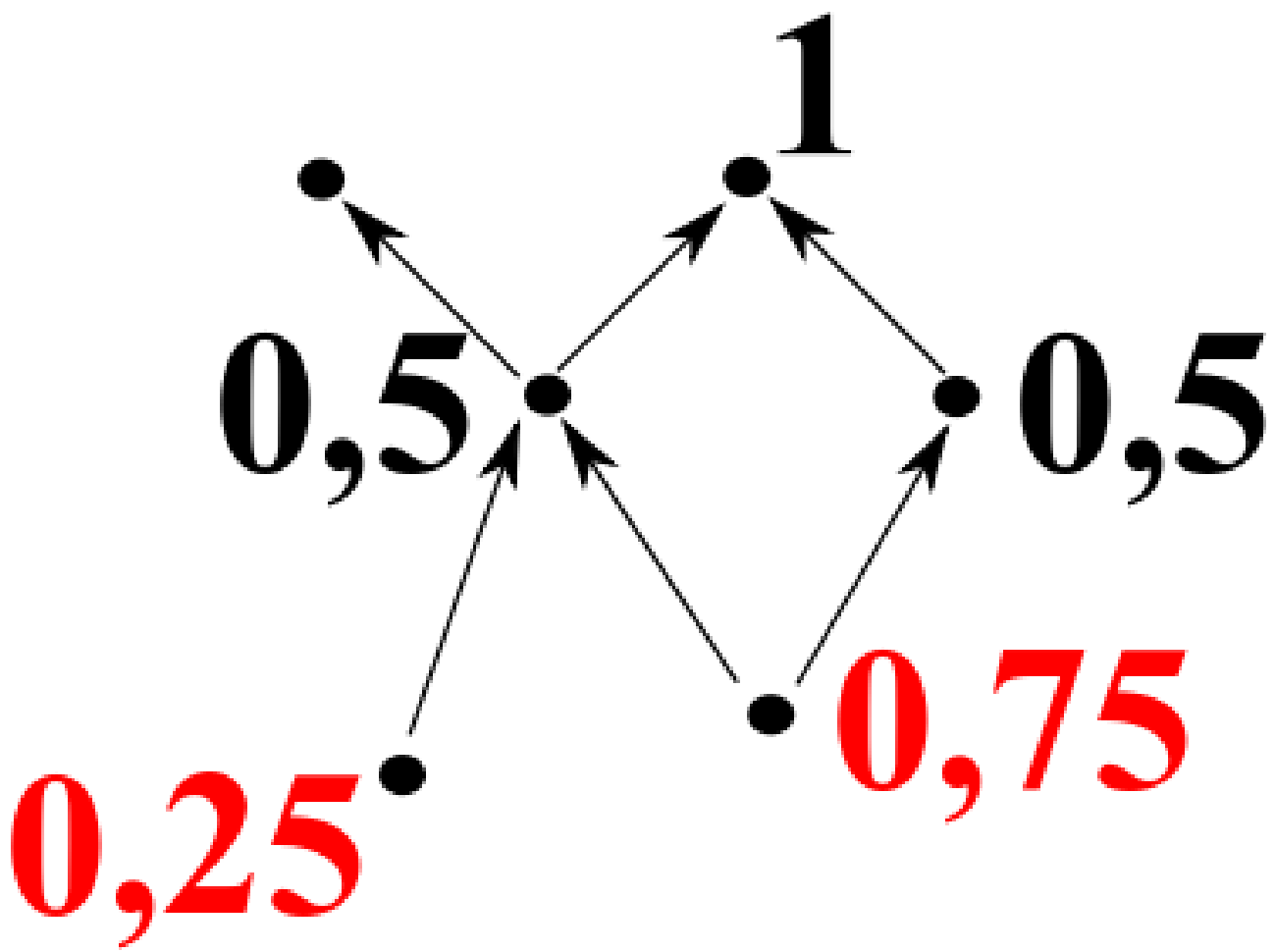
Ripartizione di punteggi



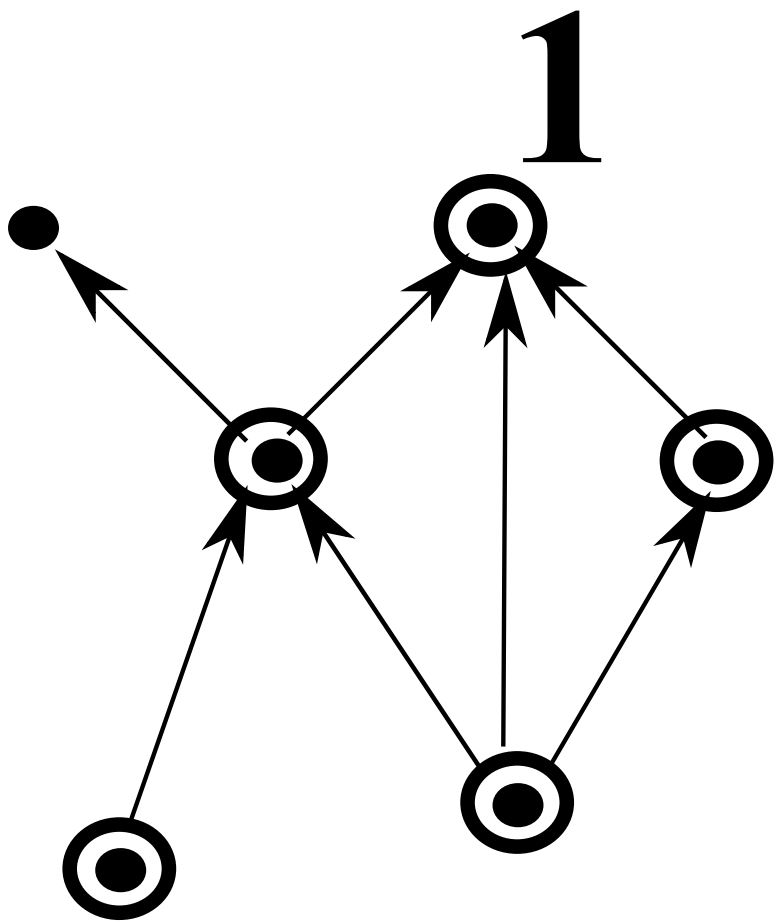
Ripartizione di punteggi



Ripartizione di punteggi

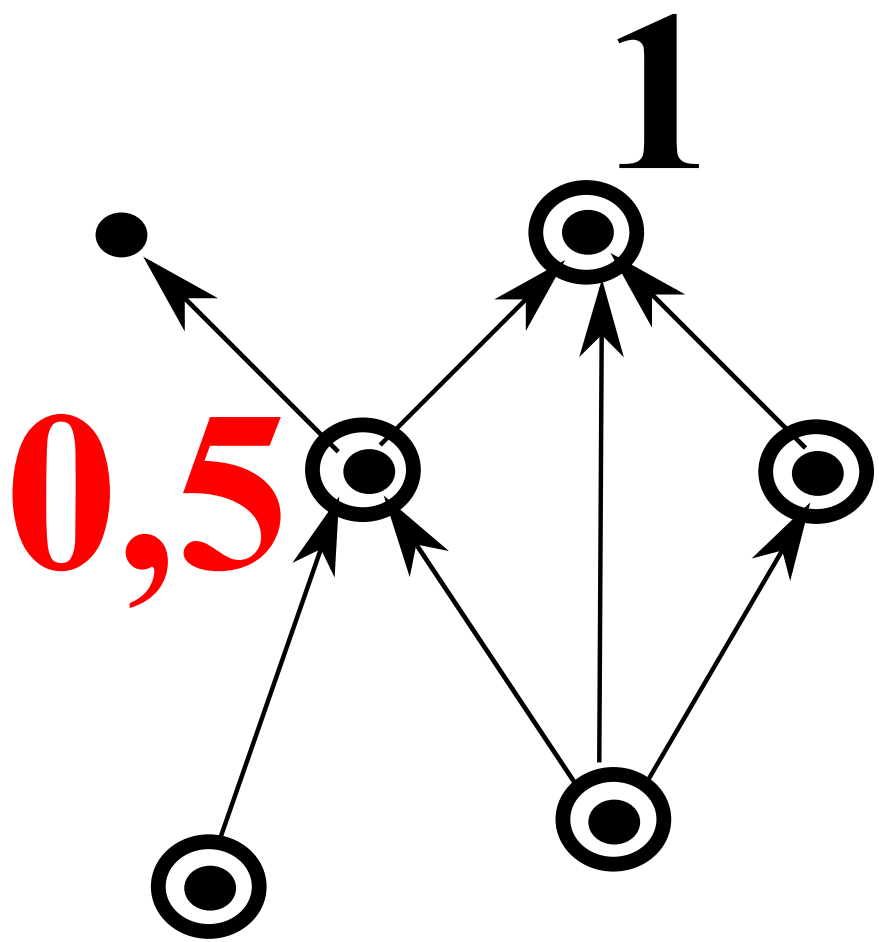


Probabilità di raggiungimento



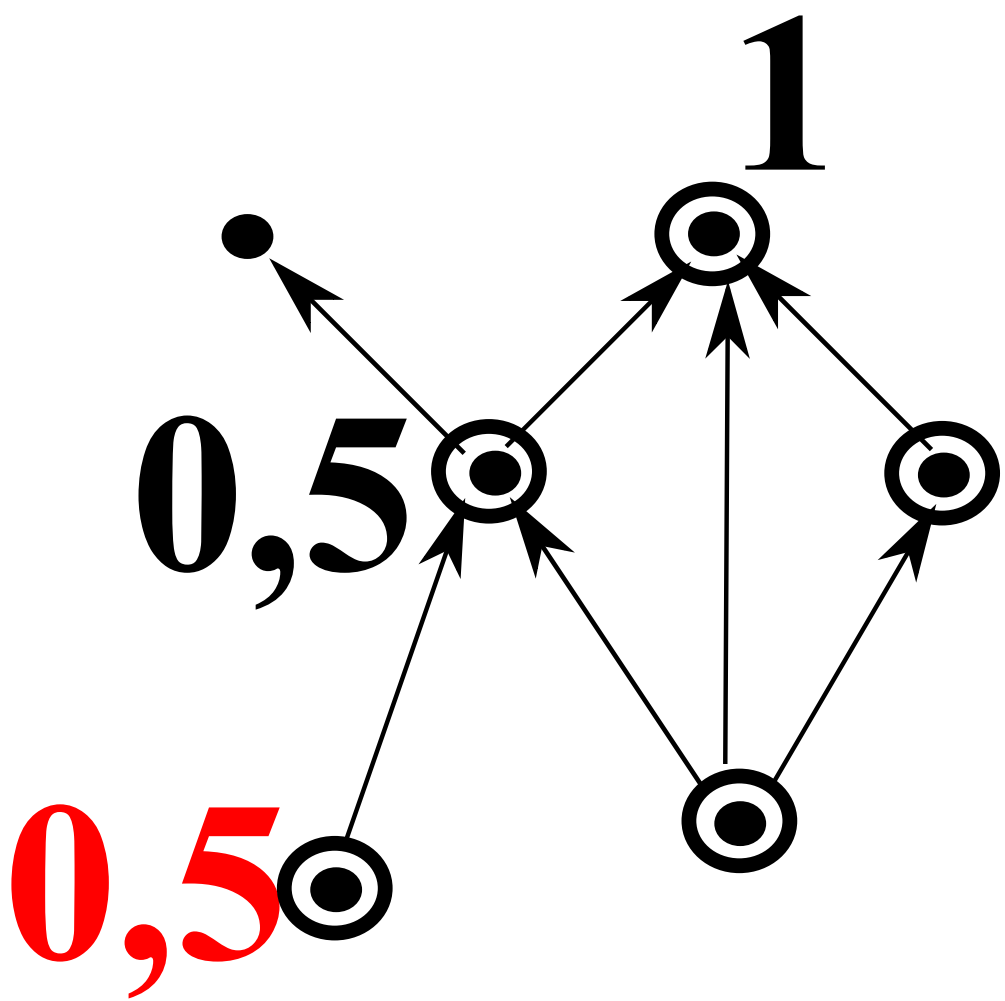
I nodi raggiungibili da una macrocategoria sono stati marcati, eventualmente da una tecnica precedente.

Probabilità di raggiungimento

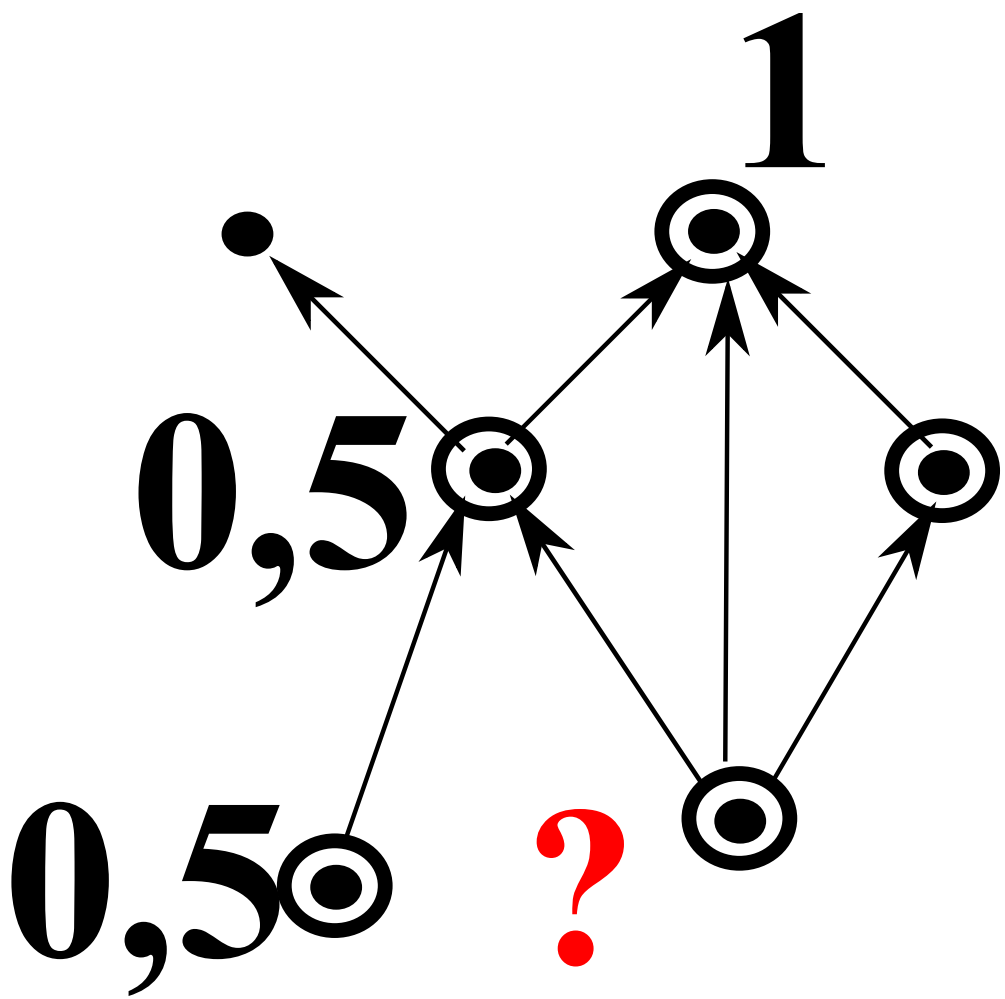


Si attraversa il grafo secondo la logica depth-first

Probabilità di raggiungimento

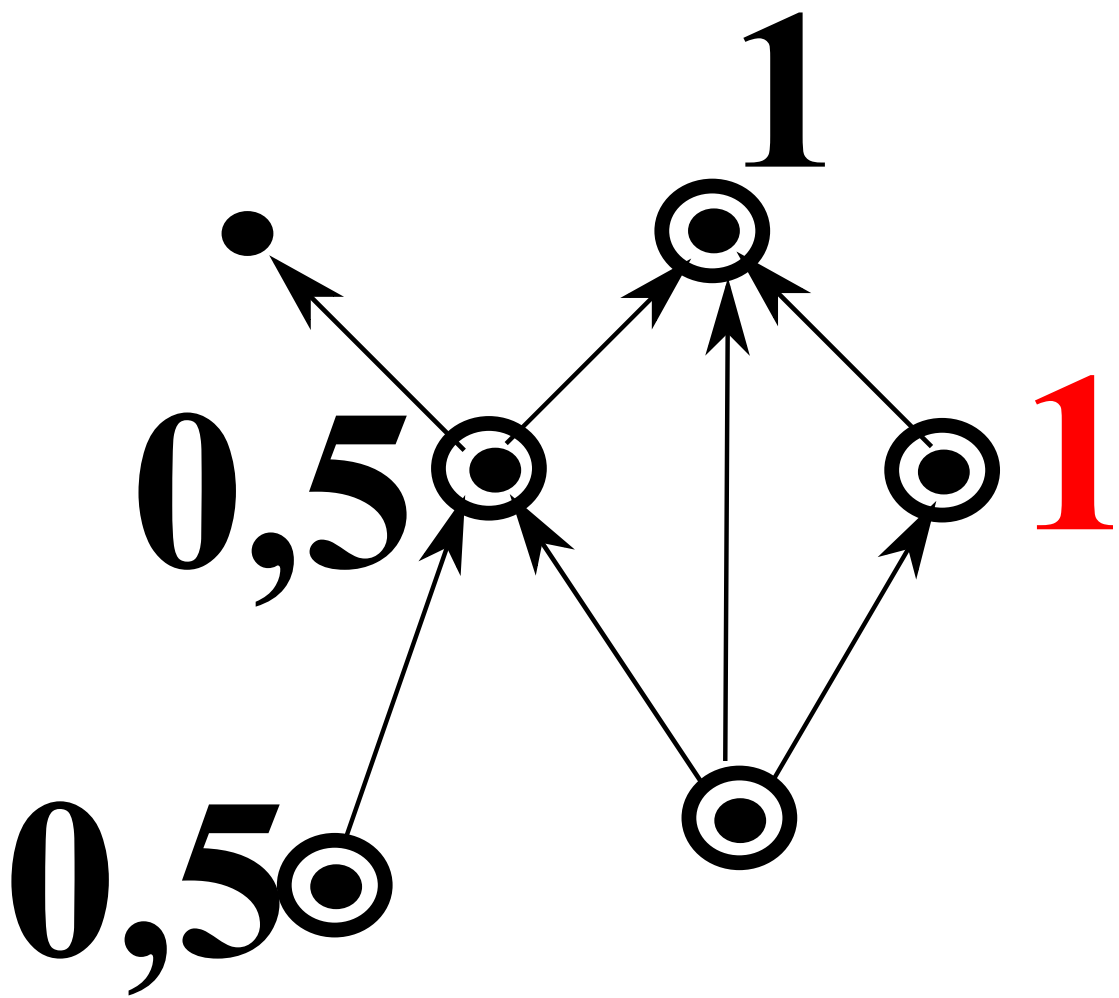


Probabilità di raggiungimento

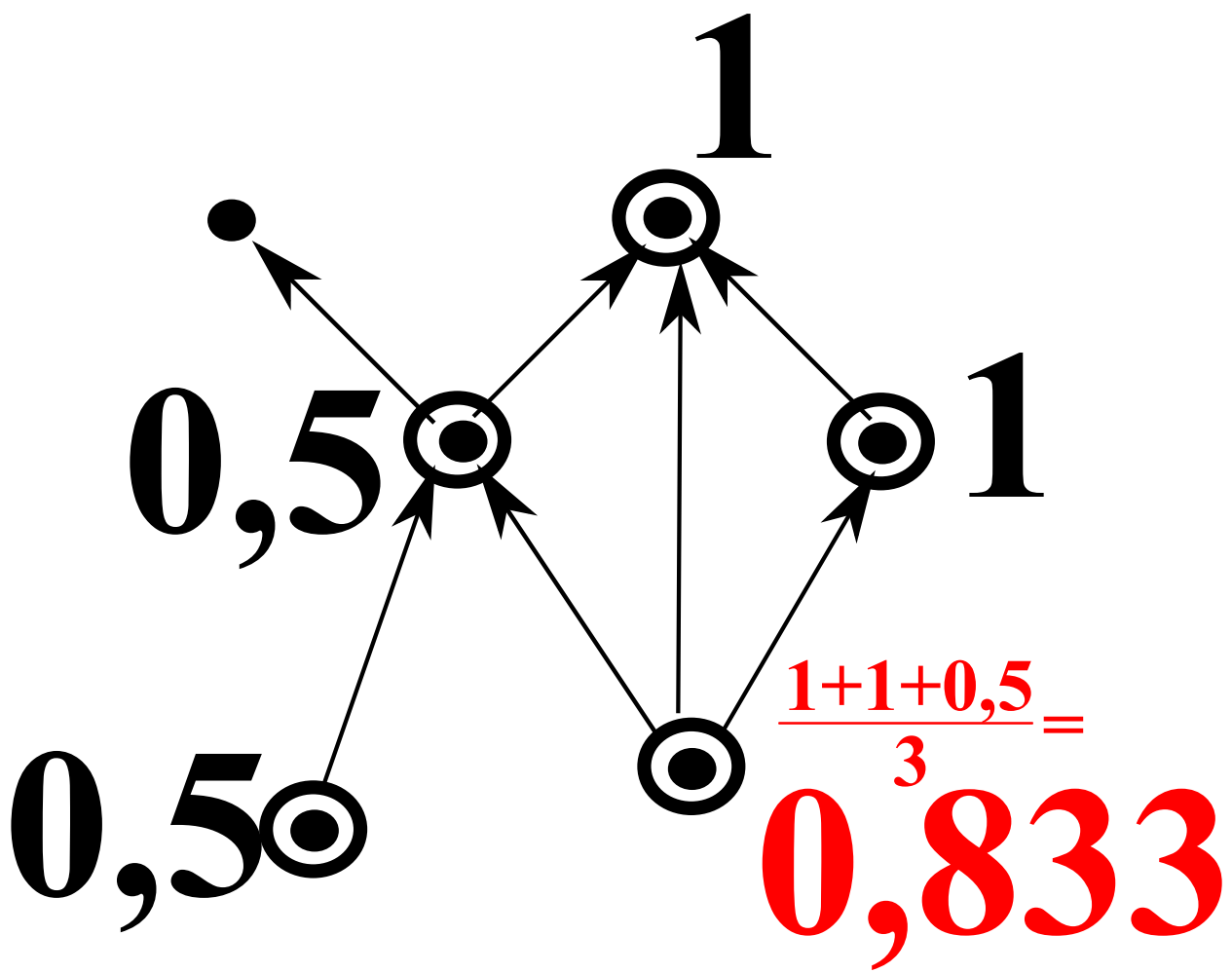


Se un nodo ha archi uscenti verso nodi marcati ma senza una probabilità assegnata, lo salto: ci passerò dopo

Probabilità di raggiungimento



Probabilità di raggiungimento



Finalmente tutti i nodi marcati sono stati elaborati, posso assenare la probabilità di raggiungimento.

Collegamenti

- Per maggiori dettagli si veda la pagina dell'AirWiki
  - http://airwiki.ws.dei.polimi.it/index.php/WikipediaCategoryGr