

Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2012/2013

**Corso di Laurea Magistrale in
Editoria e Giornalismo**

Ricerca semantica



Problema (I)

- In ogni lingua, una parola può avere più significati
- Quando si esegue una ricerca con una tale parola, i risultati più rilevanti dipendono da quale significato era inteso dall'utente
- Attualmente (fino a poche settimane fa), vengono ritornate le pagine più rilevanti per tutti i possibili significati

Problema, esempio

The image shows a screenshot of a Google search for the word "earth". The search results are displayed on a white background with a blue header. The search bar at the top shows "earth" and a magnifying glass icon. Below the search bar, there are tabs for "Web", "Images", "Maps", "Shopping", "News", "More", and "Search tools". The "Web" tab is selected. The search results are listed below, with the first result being "Google Earth".

Annotations are made on the image:

- A green oval labeled "Software?" points to the "Google Earth" result.
- A blue oval labeled "Pianeta?" points to the "Earth - Wikipedia, the free encyclopedia" result.
- An orange oval labeled "Giornata?" points to the "Earth Day Network | Earth Day 2013" result.

Search results (from top to bottom):

- Google Earth**
earth.google.com/
Google Earth lets you fly anywhere on Earth to view satellite imagery, maps, terrain, 3D buildings, from galaxies in outer space to the canyons of the ocean. Download Google Earth - Earth View in Maps - Google Планета Земля - Connect
- Earth - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Earth
Earth is the third planet from the Sun, and the densest and fifth-largest of the eight planets in the Solar System. It is also the largest of the Solar System's four ...
History of the Earth - Structure of the Earth - Age of the Earth - Earth's orbit
- Earth | Earth facts, pictures and information**
nineplanets.org/earth.html
Earth is the third planet from the Sun and the fifth largest. Find out more...
- Google Earth for mobile**
www.google.com/mobile/earth/
Features. Explore in 3D. View the same 3D imagery, terrain, and buildings* available in the desktop version of Google Earth. Local search. Search for cities ...
- NASA Earth Observatory : Home**
earthobservatory.nasa.gov/
Explore the causes and effects of climatic and environmental change through the use of real satellite data.
- Flash Earth - Zoom into satellite and aerial imagery of the Earth**
www.flashearth.com/
Satellite and aerial imagery of the Earth in Flash. Based on different data that can be chosen by a click.
- Earth**
www.universetoday.com/14367/earth/
by Fraser Cain - in 677,906 Google+ circles - More by Fraser Cain
12 Mar 2012 - The Earth is the third planet from the Sun and the fifth largest of the eight planets in our Solar System. It is located between Mars and Venus, ...
- Earth Day Network | Earth Day 2013**
www.earthday.org/
You can make a difference! Help Earth Day Network -- that works with over

5

Problema (II)

- “earth” è solo una parola, una stringa, una sequenza di caratteri
- I risultati dei diversi significati sono mescolati tra loro
 - Non sono organizzati in base al significato

6

La ricerca semantica

- Obiettivo della ricerca semantica è di
 - aggiungere semantica ai termini
 - Non più stringhe, ma concetti/entità
 - Correlare i documenti non a parole ma ai concetti
 - Organizzare i risultati in base ai concetti
 - Cercare di comprendere quale concetto era sottinteso dall'utente nella propria ricerca (ed esplicitarlo)

7

Attuale evoluzione

The image displays three screenshots of Google search results for the term "earth", illustrating the evolution of semantic search. A blue callout box labeled "Concetti" (Concepts) points to the search results in each screenshot.

Top Screenshot: A standard Google search for "earth" showing results for Google Earth, Wikipedia, and Earth Band. The "Concetti" callout points to the "Earth" entity in the Wikipedia result.

Middle Screenshot: A search for "earth" showing results for Google Earth, Wikipedia, and Earth Band. The "Concetti" callout points to the "Earth" entity in the Wikipedia result.

Bottom Screenshot: A search for "Earth" showing results for Planet Earth, Wikipedia, and Earth Band. The "Concetti" callout points to the "Earth" entity in the Wikipedia result.

Non solo google

The image displays a collage of WolframAlpha search results for the word "earth", categorized by semantic interpretation. Each category shows a different set of data and visualizations:

- Pianeta (Planet):** Shows orbital properties such as current distance from Sun (0.9835 au), largest distance from orbit center (1.52097701×10^8 km), and nearest distance from orbit center (1.47098074×10^8 km).
- Parola (Word):** Provides definitions for "earth" as a noun, including "the 3rd planet from the sun; the planet we live on" and "the loose soft material that makes up a large part of the land surface".
- Libro (Book):** Lists authors for "Earth (The Book): A Visitor's Guide to the Human Race", including Jon Stewart, David Javerbaum, Rory Albanese, Steve Bodow, Josh Lieb, Kevin Bleyer, Rich Blomquist, Tim Carvell, Wyatt Cenac, Hallie Haglund, and J. R.
- Materiale (Material):** Shows physical properties for "soil", such as angle of repose (35° for fuller's, 37.5° for moist, 37.5° for dry loam, and 50° for cracked).
- Film:** Provides basic movie information for "Earth", including title, director (Victor Asluek), runtime (33 minutes), and genres (documentary, short).
- "Mondo" (World):** Displays a world map with a satellite image link.

Altri motori di ricerca

- Esistono altri motori di ricerca semantici, anche meno conosciuti
 - Bing
 - GoPubMed
 - Motore di ricerca semantico biomedico
 - Iglue
 - Motore di ricerca semantico con capacità di annotazione/arricchimento semantico dei siti web in realtime
 - Yummly
 - Motore di ricerca semantico per ricette e cibo

Come funziona (I)

- L'idea di base è simile a quella vista per i thesauri
 - I termini riconducibili allo stesso significato e concetto vengono raggruppati in un unico insieme (Synset)
 - I synset vengono collegati tra loro in base alle relazioni semantiche esistenti tra i concetti che rappresentano
 - Si costruisce così un grafo (rete semantica) in cui
 - I nodi rappresentano i concetti/synset
 - Gli archi rappresentano le relazioni semantiche tra i concetti
 - L'informazione non strutturata (classiche pagine web) deve essere strutturata (grafo) per essere fruita al meglio
 - Vi ricorda qualcosa?

11

Come funziona (II)

- I termini ambigui di una ricerca vengono disambiguati vedendo se appartengono a synset/concetti legati tra loro
 - Termini utilizzati in una stessa ricerca, molto probabilmente, sono correlati tra loro nel significato
 - “red apple” è molto più probabile che abbia il significato di “mela rossa” che non di “Apple® rosso”

12

Google Knowledge Graph (I)

- Il nome deriva proprio dalla sua rappresentazione a grafo dei concetti
- Una ricerca
 - oltre ad essere eseguita sulle pagine web viene effettuata anche sul knowledge graph
- Quando un termine non può essere disambiguato, tutti i concetti ad esso riconducibili nel grafo vengono mostrati

13

Google Knowledge Graph, esempio (I)

The screenshot shows a Google search for 'sun'. The search bar at the top contains 'sun'. Below the search bar, the results are displayed. On the left, there are several links: 'The Sun | The Best for News, Sport, Showbiz, Celebrities | The Sun', 'Football', and 'News'. On the right, there is a Knowledge Graph card for 'The Sun'. The card includes the title 'The Sun', the location 'United Kingdom', a description 'The Sun is a daily national tabloid newspaper published in the United Kingdom and Ireland. Regionalised editions are published in Glasgow and Dublin. Wikipedia', and a 'Recent posts' section. Below the card, there is a section 'See results about' with a small image of the sun and a description 'The Sun is the star at the center of the Solar System. It is almost perfectly spherical and consists of hot ...'. A blue speech bubble points to the 'See results about' section with the text 'Cliccando su un concetto, Si inizia a “navigare” il grafo'.

Google

sun

Web Images Maps Shopping More Search tools

About 2,620,000,000 results (0.24 seconds)

[The Sun | The Best for News, Sport, Showbiz, Celebrities | The Sun](#)
[www.thesun.co.uk/](#)
Get the latest news and features at The **Sun** - Showbiz, babes, celebrities, sport and racing, national and international news. Check out the best pictures, videos, ...

[Football](#)
Get the latest football news, gossip, transfer rumours and results ...

[News](#)
Get the latest UK news, crime and political news, world news and ...

[Oracle and Sun](#)
[www.oracle.com/us/sun/index.htm](#)
Oracle announces it has entered into an agreement to acquire **Sun** Microsystems (Sun)

[The Sun NewsVoice of The Nation](#)
[sunnewsonline.com/](#)
A Nigerian newspaper with a penchant for 'British Tabloid' styled journalism. The paper and its online version lean strongly towards entertainment, politics and ...

The Sun
United Kingdom

The Sun is a daily national tabloid newspaper published in the United Kingdom and Ireland. Regionalised editions are published in Glasgow and Dublin. Wikipedia

Recent posts

17 minutes ago – The linesman who told Man City fans to celebrate with fans who'd paid £62 to see them beat Arsenal has been AXED from a game tonight.
[£62' lino axed](#)

Feedback / More info

See results about

 **Sun**
The Sun is the star at the center of the Solar System. It is almost perfectly spherical and consists of hot ...

14

Google Knowledge Graph (II)

- La struttura a grafo permette di navigare tra i concetti
 - Ci si sposta da un nodo/synset/concetto ad altro ad esso collegato
 - La ricerca (Web e/o su KG) corrispondente viene eseguita



Google Knowledge Graph, esempio (II)

Google search results for "sun life cycle".

Search results include:

- [Sun - Wikipedia, the free encyclopedia](#)
The **Sun** by the Atmospheric Imaging Assembly of NASA's Solar Dynamics Observatory - 20100819.jpg **Life-cycle** of the **Sun**; sizes are not drawn to scale.
[Sun \(disambiguation\)](#) - [Sun path](#) - [Faint young Sun paradox](#) - [Sun dog](#)
- [Life cycle of our Sun - YouTube](#)
www.youtube.com/watch?v=5oCh4XTQIV4
Mar 6, 2011 - Uploaded by sands0sands
A description of the life **cycle** of our Sun and other G - type stars.

More videos for **sun life cycle** »

[Sun Lifecycle](#)
cde.nwc.edu/SCI2108/course.../the_sun/sun_lifecycle/lifecycle.htm
The **lifecycle** of any star is determined by the mass of the star. Our **Sun** is an average-mass object, and it lives an average **life**. Stars of lower mass live very ...

Sun

The Sun is the star at the center of the Solar System. It is almost perfectly spherical and consists of hot plasma interwoven with magnetic fields. [Wikipedia](#)

Surface temperature: 5,778 K
Radius: 1 R_☉
Mass: 1.989E30 kg
Distance to Earth: 92,960,000 miles (149,600,000 km)
Coordinates: RA 19h 4m 31s | Dec 63° 52.200'
Orbits: Galactic Center

People also search for

[Moon](#) [Earth](#) [Mars](#) [Jupiter](#) [Venus](#)

Cliccando su un altro concetto, lo si "esplora"

Google Knowledge Graph, esempio (III)

The screenshot shows a Google search for "earth's moon". The search bar is at the top with the Google logo. Below the search bar, there are tabs for "Web", "Images", "Maps", "Shopping", "News", "More", and "Search tools". The "Web" tab is selected. Below the tabs, there is a row of images representing celestial bodies: Sun, Moon, Earth, Mars, Jupiter, Venus, Mercury, Saturn, Neptune, and Uranus. The "Moon" image is highlighted with a red box. To the left of the "Moon" image, there is a small box with the text "People also search for". Below the row of images, there are several search results. The first result is "Solar System Exploration: Planets: **Earth's Moon**. Overview" from solarsystem.nasa.gov. The second result is "Moon - Wikipedia, the free encyclopedia" with a snippet about synchronous rotation. The third result is "Moon - Facts and Information about the **Earth's Moon** | Space.com" with a snippet about the formation of the Moon. To the right of the search results, there is a large image of the Moon and a smaller image showing the phases of the Moon. Below the large image, there is a section titled "Moon" with a description: "The Moon is the only natural satellite of the Earth, and the fifth largest satellite in the Solar System. Wikipedia". Below the description, there is a line of text: "Distance to Earth: 238,900 miles (384,400 km)".

17

Google Knowledge Graph (III)

- Knowledge Graph ha attualmente affiancato e integrato la “classica” ricerca web
- Si può
 - Navigare nel grafo
 - Navigare i risultati della ricerca testuale web
- Inoltre eventuali termini “base” nella ricerca effettuata sul KG vengono ricondotti ad apposite ricerche web classiche, prestabilite anche dai termini che identificano il concetto “in cui ci si trova”

18

Come viene costruito il KG? (I)

- Informazioni sui concetti vengono tratte da database pubblici, inclusi:
 - Freebase
 - database gratuito e aperto di oltre 24 milioni di cose, tra cui film, libri, spettacoli televisivi, personaggi famosi, luoghi, aziende
 - Wikipedia
 - CIA World Factbook
 - pubblicazione annuale della CIA che riporta i dati statistici fondamentali e una sintesi di informazioni riguardanti tutti i paesi del mondo
 - Risorse specifiche per argomento
 - e.g., Weather Underground per informazioni meteo e la Banca Mondiale per statistiche economiche
 - I dati di ricerca di Google
 - usati per misurare la popolarità di un argomento e contribuire a decidere quali sono le informazioni più richieste

19

Come viene costruito il KG? (II)

- I concetti e le loro relazioni vengono inferiti
 - Utilizzando thesauri e ontologie già esistenti
 - Automaticamente analizzando i termini chiave delle pagine web e come le pagine web sono collegate dai link
- Secondo gli ultimi dati disponibili, nel 2012 Knowledge Graph contiene dati su
 - 570M di oggetti
 - 18G di fatti

20

Google Knowledge Graph (IV)

- Il motto del progetto è “Things, not strings” (“cose, non parole”)
 - Il focus delle ricerche si sposta dai termini/keyword agli oggetti/entità che esse rappresentano
- Google: non più “motore di ricerca”, ma “motore di conoscenza”

21

Conseguenze sul SEO

- Anche le tecniche di SEO risentono dello cambiamento di focus
- Attualmente Google indicizza nel KG solo alcune informazioni delle pagine web
 - persone, prodotti, aziende, ricette, eventi, musica, video
 - Questa informazione deve essere strutturata
 - microdata

22

Microdata

- Estensione di HTML
- Permette di annotare/arricchire le pagine web (HTML) con nuovi marcatori che includono informazioni semantiche

23

Microdata, esempio (I)

```
<section> Hello, my name is John Doe, I am a graduate research  
assistant at the University of Dreams. My friends call me Johnny.  
You can visit my homepage at <a  
href="http://www.JohnnyD.com">www.JohnnyD.com</a>.  
I live at 1234 Peach Drive Warner Robins, Georgia.</section>
```



```
<section itemscope itemtype="http://schema.org/Person">  
  Hello, my name is <span itemprop="name">John Doe</span>,  
  I am a <span itemprop="jobTitle">graduate research assistant</span>  
  at the <span itemprop="affiliation">University of Dreams</span>.  
  My friends call me <span itemprop="additionalName">Johnny</span>.  
  You can visit my homepage at  
  <a href="http://www.JohnnyD.com" itemprop="url">www.JohnnyD.com</a>.  
  <section itemprop="address" itemscope  
    itemtype="http://schema.org/PostalAddress">  
    I live at <span itemprop="streetAddress">1234 Peach Drive</span>  
    <span itemprop="addressLocality">Warner Robins</span>,  
    <span itemprop="addressRegion">Georgia</span>.  
  </section>  
</section>
```

24

Microdata, esempio (I)

Un web engine comprenderà il microdata come

Item

```
Type: http://schema.org/Person
name = John Doe
jobTitle = graduate research assistant
affiliation = University of Dreams
additionalName = Johnny
url = http://www.johnnyd.com/
address = Item(1)
```

Item 1

```
Type: http://schema.org/PostalAddress
streetAddress = 1234 Peach Drive
addressLocality = Warner Robins
addressRegion = Georgia
```

25

Semantica di microdata

- Gli oggetti, con i loro attributi, definibili in microdata sono definiti in <http://schema.org>

Thing: additionalType, description, image, name, url

CreativeWork: about, accountablePerson, aggregateRating, alternativeHeadline, associatedMedia, audience, audio, author, award, awards, comment, contentLocation, contentRating, contributor, copyrightHolder, copyrightYear, creator, dateCreated, dateModified, datePublished, discussionUrl, editor, encoding, encodings, genre, headline, inLanguage, interactionCount, isFamilyFriendly, keywords, mentions, offers, provider, publisher, publishingPrinciples, review, reviews, sourceOrganization, text, thumbnailUrl, version, video

Article: articleBody, articleSection, wordCount

BlogPosting

NewsArticle: dateline, printColumn, printEdition, printPage, printSection

ScholarlyArticle

MedicalScholarlyArticle: citation, publicationType

Blog: blogPost, blogPosts

Book: bookEdition, bookFormat, illustrator, isbn, numberOfPages

Comment

Diet: dietFeatures, endorers, expertConsiderations, overview, physiologicalBenefits, proprietaryName, risks

ExercisePlan: activityDuration, activityFrequency, additionalVariable, exerciseType, intensity, repetitions, restPeriods, workload

ItemList: itemListElement, itemListOrder

Map

MediaObject: associatedArticle, bitrate, contentSize, contentUrl, duration, embedUrl, encodesCreativeWork, encodingFormat, expires, height, interactionCount, offers, playerType, regionsAllowed, requiresSubscription, uploadDate, width

AudioObject: transcript

ImageObject: caption, exifData, representativeOfPage, thumbnail

MusicVideoObject

VideoObject: caption, productionCompany, thumbnail, transcript, videoFrameSize, videoQuality

Movie: actor, actors, director, duration, musicBy, producer, productionCompany, trailer

MusicPlaylist: numTracks, track, tracks

MusicAlbum: byArtist

MusicRecording: byArtist, duration, inAlbum, inPlaylist

Painting

Photograph

Recipe: cookingMethod, cookTime, ingredients, nutrition, prepTime, recipeCategory, recipeCuisine, recipeInstructions, recipeYield

26

Riferimenti

- Ricerca semantica
 - http://en.wikipedia.org/wiki/Semantic_search
 - <http://searchenginewatch.com/>
- Google Knowledge Graph
 - <http://www.google.com/intl/it/insidesearch/features/search/knowledge.html>
 - <http://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html>
 - <http://support.google.com/websearch/bin/answer.py?hl=it&answer=2620861>
 - <http://searchenginewatch.com/search?q=%22knowledge+graph%22>
- Ricerca semantica e SEO
 - <http://searchenginewatch.com/article/2214849/Googles-Knowledge-Graph-Implications-for-Search-SEO>
 - <http://searchenginewatch.com/article/2234448/Keywords-Are-Dead-Long-Live-User-Intent>