

Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2013/2014

**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 (stato in evoluzione), vengono ritornate le pagine più rilevanti per tutti i possibili significati

Problema, esempio

The image shows a Google search results page for the query 'earth'. The search bar at the top shows 'earth' and the search button. Below the search bar, there are tabs for 'Web', 'Images', 'Maps', 'Shopping', 'News', 'More', and 'Search tools'. The search results are displayed below, with a green box labeled 'Software?' pointing to the 'Google Earth' result, a blue box labeled 'Pianeta?' pointing to the 'Earth - Wikipedia, the free encyclopedia' result, and an orange box labeled 'Giornata?' pointing to the 'Earth Day Network | Earth Day 2013' result. The search results include links to Google Earth, Wikipedia, NASA Earth Observatory, Flash Earth, and Earth Day Network.

Google
earth

About 1,220,000,000 results (0.24 seconds)

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 Платформа 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 i...
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 screenshot shows a Google search for 'earth'. The results are categorized into three main semantic entities, each with a green arrow pointing to a 'Concetti' (Concepts) callout box:

- Google Earth**: A virtual globe, map and geographical information program that was originally ...
- Earth (planet)**: Earth is the third planet from the Sun, and the densest and fifth-largest of the eight planets in the Solar System ...
- Earth (band)**: Earth is an American musical group based in Seattle, Washington, formed in 1989 and led by the guitarist Dylan Carlson. Earth's music is nearly all instrumental ...

The 'Concetti' callout box is a speech bubble containing the word 'Concetti'.

Non solo google

The collage shows six different WolframAlpha search results for the word "earth", each with a different interpretation highlighted in a blue box:

- Pianeta**: Shows orbital properties of Earth as a planet, such as current distance from Sun (0.9835 au) and largest distance from orbit center (1.52097701×10^8 km).
- Parola**: Shows definitions of "earth" as a word, including "the 3rd planet from the sun" and "the loose soft material that makes up a large part of the land surface".
- Libro**: Shows information about the book "Earth (The Book): A Visitor's Guide to the Human Race" by Jon Stewart, David Javerbaum, and Rory Albanese.
- Materiale**: Shows physical properties of soil, such as angle of repose (35° for Fuller's) and moisture content (37.5% for moist soil).
- Film**: Shows basic movie information for the film "Earth" directed by Victor Aslruk, with a runtime of 33 minutes.
- "Mondo"**: Shows a world map and location information for "world".

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 (le più disparate) 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)

- Si entra nell'ambito della “rappresentazione della conoscenza” e dei “sistemi esperti”
- In questo campo non si parla di “thesauro” ma di “ontologia” e “knowledge base”

Come funziona (III)

- 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”

13

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

14

Google Knowledge Graph, esempio (I)

The screenshot shows a Google search for "sun". The search bar at the top contains the word "sun". Below the search bar, there are tabs for "Web", "Images", "Maps", "Shopping", "More", and "Search tools". The "Web" tab is selected. The search results show "About 2,620,000,000 results (0.24 seconds)". The first result is "The Sun | The Best for News, Sport, Showbiz, Celebrities | The Sun" with the URL "www.thesun.co.uk/". Below this, there are links for "Football" and "News". A blue speech bubble points to the "News" link with the text "Cliccando su un concetto, Si inizia a 'navigare' il grafo". To the right of the search results, there is a Knowledge Graph entry for "The Sun". It 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 Knowledge Graph entry, 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 ...".

15

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".

Web Images Maps Shopping Videos More Search tools

About 25,900,000 results (0.34 seconds)

Sun - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Sun
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=5oCh4XTQIVA
Mar 6, 2011 - Uploaded by sands0sands
A description of the life cycle of our Sun and other G - type stars.

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"

17

Google Knowledge Graph, esempio (III)

Google search results for "earth's moon".

Web Images Maps Shopping News More Search tools

People also search for
Moon Earth Mars Jupiter Venus Mercury Saturn Neptune Uranus

Solar System Exploration: Planets: Earth's Moon: Overview
solarsystem.nasa.gov > Planets
13 Dec 2012 - Our Moon makes Earth a more livable planet by moderating our home planet's wobble on its axis, leading to a relatively stable climate, and ...

Moon - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Moon
Jump to **Appearance from Earth**: The Moon is in synchronous rotation: it rotates about its axis in about the same time it takes to orbit the Earth. This results ...
Orbit of the Moon - Moon landing - Moon-landing conspiracy theory - Natural satellite

Moon - Facts and Information about the Earth's Moon | Space.com
www.space.com/55-earths-moon-formation-composition-and-orbit.ht...
Scientists say a giant impact knocked off the raw ingredients for the moon off the primitive molten Earth and into orbit. SPACE.com has an overview of Earth's ...

Earth's Moon Phases: Monthly Lunar Cycles (Infographic) | Space.com

Moon
The Moon is the only natural satellite of the Earth, and the fifth largest satellite in the Solar System. Wikipedia

Distance to Earth: 238,900 miles (384,400 km)

18

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”

19

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

20

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 alcuni dati disponibili, nel 2012 Knowledge Graph contiene dati su
 - 570M di oggetti
 - 18G di fatti

21

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”

22

Conseguenze sul SEO

- Anche le tecniche di SEO risentono del 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

23

Microdata

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

24

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>
```

25

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
```

26

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

27

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>

28