Automatic Meaning Discovery Using Google

Paul Vitanyi
CWI, University of Amsterdam, National ICT Australia

Joint work with Rudi Cilibrasi

New Scientist, Jan. 29, 2005

New Scientist Google's search for meaning - Technology http://www.newscientist.com/channel/i

INFO-TECH

Google's search for meaning

29 January 2005
From New Scientist Print Edition. Subscribe and get 4 free issues.

Duncan Graham-Rowe

COMPUTERS can learn the meaning of words simply by plugging into Google. The finding could bring forward the day that true artificial intelligence is developed.

Trying to get a computer to work out what words mean - distinguish between "rider" and "horse" say, and work out how they relate to each other - is a long-standing problem in artificial intelligence research.

One of the difficulties has been working out how to represent knowledge in ways that allow computers to use it. But suddenly that is not a problem any more, thanks to the massive body of text that is available, ready indexed, on search engines like Google (which has more than 8 billion pages indexed).

The meaning of a word can usually be gleaned from the words used around it. Take the word "rider". Its meaning can be deduced from the fact that it is often found close to words like "horse" and "saddle". Rival attempts to deduce meaning by relating hundreds of thousands of words to each other require the creation of vast, elaborate databases that are taking an enormous amount of work to construct.

But Paul Vitanyi and Rudi Cilibrasi of the National Institute for Mathematics and Computer Science in Amsterdam, the Netherlands, realised that a Google search can be used to measure how closely two words relate to each other. For instance, imagine a computer needs to understand what a hat is.

To do this, it needs to build a word tree - a database of how words relate to each other. It might start off with any two words to see how they relate to each other. For example, if it googles "hat" and "head" together it gets nearly 9 million hits, compared to, say, fewer than half a million hits for "hat" and "banana". Clearly "hat" and "head" are more closely related than "hat" and "banana".

To gauge just how closely, Vitanyi and Cilibrasi have developed a statistical indicator based on these hit counts that gives a measure of a logical distance separating a pair of words. They call this the normalised Google distance, or NGD. The lower the NGD, the more closely the words are related.

" The web might make all the difference to whether we make an artificial intelligence or not"

By repeating this process for lots of pairs of words, it is possible to build a map of their distances, indicating how closely related the meanings of the words are. From this a computer can infer meaning, says Vitanyi. "This is automatic meaning extraction. It could well be the way to make a computer understand things and act semi-intelligently," he says.

The technique has managed to distinguish between colours, numbers, different religions and Dutch painters based on the number of hits they return, the researchers report in an online preprint (www.arxiv.org/abs/cs.CL/0412098).

The pair's results do not surprise Michael Witbrock of the Cyc project in Austin, Texas, a 20-year effort to create an encyclopaedic knowledge base for use by a future artificial intelligence. Cyc represents a vast quantity of fundamental human knowledge, including word meanings, facts and rules of thumb. Witbrock believes the web will ultimately make it possible for computers to acquire a very detailed knowledge base. Indeed, Cyc has already started to draw upon the web for its knowledge. "The web might make all the difference in whether we make an artificial intelligence or not," says Witbrock.

From issue 2484 of New Scientist magazine, 29 January 2005, page 21

More Info-Tech Stories

- On the trail of the zombie PCs
- The personalised traffic jam buster
- Introducing the glooper computer
- School computers may be overrated
- New Powerbook controlled with a shake

More Stories

Explore: info-tech

Printable version | Send to a friend | RSS feed XML

For exclusive news and expert analysis every week subscribe to New Scientist Print Edition

For whats in New Scientist magazine this week see contents

Search all stories

Contact us about this story

Sign up for our free newsletter

Slashdot: News for Nerds; Stuff that Matters, Jan. 28, 2005

Slashdot
News for Nerds. Stuff that matters.

Login
Why Login?
Why Subscribe?

Sections
Main
Apache
Apple
2 more
AskSlashdot
5 more
Books
BSD
Developers
3 more
Games
12 more
Interviews
IT
7 more
Linux
3 more
Politics
Science
YRO
1 more

Help
FAQ
Bugs

Stories
Old Stories
Old Polls
Topics
Hall of Fame
Submit Story

About
Supporters
Code
Awards

Services
Broadband
PriceGrabber
Product Guide
Special Offers
Tech Jobs

Deriving Semantic Meaning From Google Results

Posted by michael on Sat Jan 29, '05 04:35 PM
from the can-also-use-tea-leaves-if-google-not-available dept.

prostoalex writes *"New Scientist talks about Paul Vitanyi and Rudi Cilibrasi of the National Institute for Mathematics and Computer Science in Amsterdam and their work to extract meaning of words from Google's index. The pair demonstrates an unsupervised clustering algorithm, which 'distinguish between colours, numbers, different religions and Dutch painters based on the number of hits they return', according to New Scientist."*

Slashdot Log In
Nickname:
Password:
Public Terminal
Log in
[Create a new account]

Related Links
- Compare prices on Scientific Products
- Compare prices on Google Related Items
- Review IT Products
- Review Google Products
- prostoalex
- talks about
- their work to extract meaning of words from Google's index
- More Google stories
- More Science stories

This discussion has been archived. No new comments can be posted.

Deriving Semantic Meaning From Google Results | Log in/Create an Account | Top | **120** comments | Search Discussion

Threshold: -1: 120 comments | Threaded | Oldest First | Change

The Fine Print: The following comments are owned by whoever posted them. We are not responsible for them in any way.

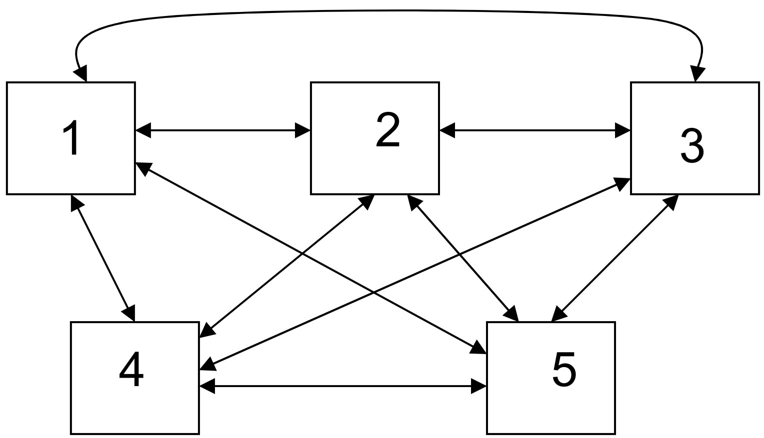
The elephant in the living room. (Score:4, Insightful)
by Eunuch (844280) * on Saturday January 29, @04:36PM (#11515623)

These kinds of articles never seem to get a very basic problem--natural languages. English is full of words that trip even humans. "Right" the direction versus "right" the judgement is a good example. In wartime something as simple as that may have lead to

Dutch Radio: TROS Radio Online, March 8, 2005

The Problem:

Given: Literal objects (binary files)



Determine: "Similarity" Distance Matrix (distances between every pair)

Applications: Clustering, Classification, Evolutionary trees of Internet documents, computer programs, chain letters, genomes, languages, texts, music pieces, ocr, ……

TOOL:

- Information Distance (Li, Vitanyi, 96; Bennett,Gacs,Li,Vitanyi,Zurek, 98)

$D(x,y) = \min \{ |p|: p(x)=y \ \&\ p(y)=x\}$

Binary program for a Universal Computer
(Lisp, Java, C, Universal Turing Machine)

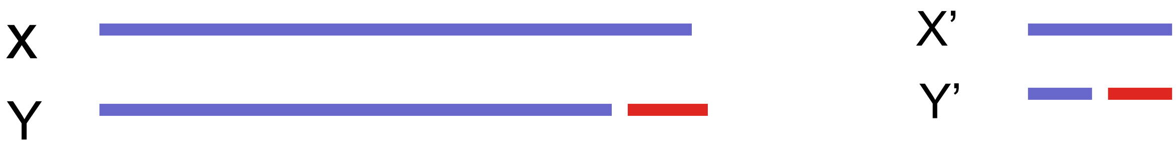
Theorem (i) $D(x,y) = \max \{K(x|y),K(y|x)\}$

Kolmogorov complexity of x given y, defined as length of shortest binary ptogram that outputs x on input y.

(ii) $D(x,y) \leq D'(x,y)$ Any computable distance satisfying $\sum_y 2^{-D'(x,y)} \leq 1$ for every x.

(iii) $D(x,y)$ is a metric.

However:

- 

  D(x,y)=D(x',y') = But x and y are much more similar than x' and y'

- So, we Normalize:

  Li Badger Chen Kwong Kearney Zhang 01
  Li Vitanyi 01/02
  Li Chen Li Ma Vitanyi 04

- $$d(x,y) = \frac{D(x,y)}{\mathrm{Max}\,\{K(x),K(y)\}}$$

  Normalized Information Distance (NID)
  The "Similarity metric"

Properties NID:

- Theorem: (i) $0 \le d(x,y) \le 1$
  (ii) $d(x,y)$ is a metric
  symmetric,triangle
  inequality, $d(x,x)=0$
  (iii) $d(x,y)$ is universal
  $d(x,y) \le d'(x,y)$ for every
  computable, normalized
  ($0 \le d'(x,y) \le 1$) distance
  satisfying standard “density”
  condition.
- Drawback: $NID(x,y) = d(x,y)$ is noncomputable, since $K(.)$ is!

In Practice:

- Replace NID(x,y) by

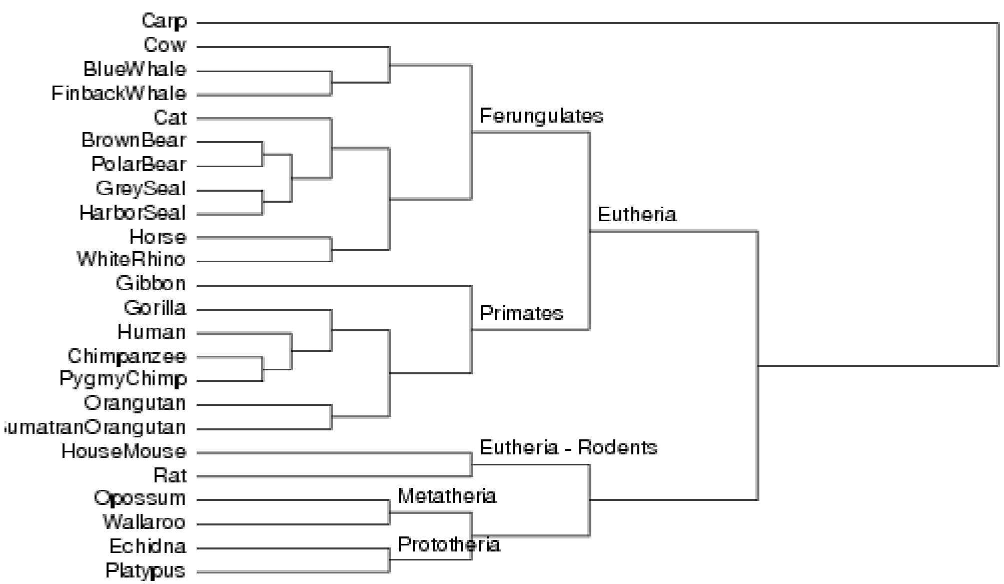
$$NCD(x,y) = \frac{C(xy) - \min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}$$

Normalized Compression Distance (NCD)

Length (#bits) compressed version x using compressor C (gzip, bzip2, PPMZ,…)

- This NCD is the same formula as NID, but rewritten using “C” instead of “K”

Example:
Phylogeny using whole mtDNA of species



Genomics just one example;
also used with (e.g.):
MIDI music files (music clustering)
Languages (language tree)
Literature
OCR

Time sequences:
(All data bases used in all major data-mining conferences of last 10Y)
Superior over all methods:
In: Anomaly detection
Heterogenous data

New Scientist, March 28, 2003

Software to unzip identity of unknown composers

19:00 09 April 2003
Exclusive from New Scientist Print Edition
Hazel Muir

A standard PC file-compression program can tell the difference between classical music, jazz and rock, all without playing a single note. This new-found ability could help scholars identify the composers of music that until now has remained anonymous.

The technique exploits the ability of off-the-shelf "zip" data-compression software to do more than just squeeze PC files into manageable sizes. For instance, various zip programs have already been used to detect the language a piece of text is written in (New Scientist print edition, 15 December 2001).

To do this, you first take several long text files, each in a known language, and compress them, noting the file size of each. You then append the unknown file to each of the uncompressed, known files in turn, and compress them again, noting the difference that adding the unknown file makes in each case.

The smaller the difference, the more likely the languages are to be the same. That is because the zip program looks for duplicated sequences in the text to shrink it without losing information.

Rudi Cilibrasi, Paul Vitanyi and Ronald de Wolf of the Dutch National Research Institute in Amsterdam wondered if such compression could also help distinguish between musical genres. So they tried it out on digital files of various pieces, including some from Beethoven, Miles Davis and Jimi Hendrix.

Rhythm and melody

They subtracted any data unrelated to the actual music, such as digital ID tags, to create a data string representing only the rhythm and melody of the tune. Using a program called Bzip2, they followed a similar procedure as with the text files, measuring how similar each piece was to every other. Then they plotted the results in a way that produces a tree-shaped pattern, in which similar pieces cluster together on the same branch.

In a test with 12 each of jazz, classical and rock pieces, the results were fairly good. Ten of the jazz, nine of the rock and most of the classical pieces ended up in three distinct branches of the tree.

When applied to 32 classical pieces, the technique clustered each composer on a separate branch. Vitanyi thinks the trick could help identify a plausible composer for works of unknown origin, as long as they have written several known works for comparison. It could also help online music stores, for example by classifying music files.

The technique's elegance lies in the fact that it is tone deaf. Rather than looking for features such as common rhythms or harmonies, says Vitanyi, "it simply compresses the files obliviously."

"I would love a technique that can work out who wrote something just by putting the notes on a page into a computer," says Jeremy Summerly of the Royal Academy of Music in London, who tries to identify the composers of unattributed fragments of 16th-century musical scores. The technique is promising, he says, because it detects features of a piece that the composer does not consciously think about, but which are actually their hallmark.

Summerly hopes to see what the technique makes of the second half of Mozart's Requiem, completed by Franz Süssmayr after Mozart's death. The way it clusters among other works by Mozart and Süssmayr might reveal how much original work Süssmayr contributed.

Related Articles

- Insect swarming inspires jazz software
  6 May 2002
- Computer DJ uses biofeedback to pick tracks
  14 November 2001
- 'Classical karaoke' is helping budding maestros conduct an orchestra
  28 March 2001

Search New Scientist
Contact us

Web Links

- Dutch National Research Institute
- Royal Academy of Music
- How file compression works

Izvestia, April 2003

…ожно предположить,
…ч лет назад людям уже был известен годичный
…тила.

…в Ирландии надгробные сооружения во время летнего и зимнего солнцестояния обращены к восходящему Солнцу, которое достигает в это время своей самой северной или самой южной точки на небосводе, что легко увидеть в любом месте земного шара.

АРХИВАТОР РАЗБИРАЕТСЯ В МУЗЫКЕ

С помощью обычной компьютерной программы сжатия файлов можно отличить классическую музыку от джаза и рока, не воспроизводя ни одной ноты. Это неожиданное открытие поможет определить, кому из композиторов принадлежат музыкальные произведения, авторы которых до сих пор считались неизвестными.

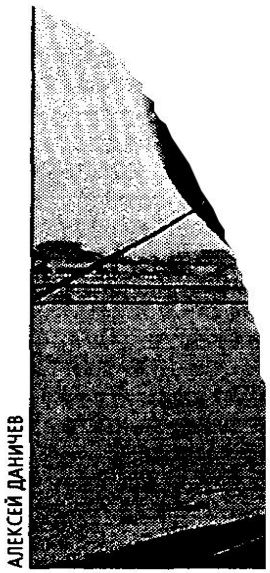
Программы архивирования данных, наподобие zip, не только сжимают файлы в приемлемые по размерам архивы. Их можно использовать и для распознавания языка, на котором написан отрывок текста.

Руди Цилибраси, Пол Витыньи и Рональд де Вольф из Голландского национального исследовательского института в Амстердаме решили посмотреть, можно ли использовать эту особенность, чтобы распознавать музыку разных жанров. Отличительная особенность приема в том, что для различения музыкальных жанров не требуется проигрывать ни одной ноты. Вместо того чтобы искать общие мелодические и ритмические рисунки, программа просто сжимает звуковые файлы.

ЗАПАХ ПОРАЖЕНИЯ

Запах влияет на наше поведение, так как мы склонны запоминать его в связи с эмоциями, которые ему сопутствовали, считает психолог из университета Брауна (США).

Рейчел Херц провела серию экспериментов с участием студенток университета Брауна, чтобы проверить, как влияет запах на эмоции и по… Шестидесяти трем участницам исследования было предло… …ую игру, победить в которой…



Если атомоход д…

Дер…
пр…

Слу…

CompLearn Toolkit

- home |
- documentation |
- download |
- forums |
- license |
- development
- **What is CompLearn?**
- The CompLearn Toolkit is a suite of simple-to-use utilities that you can use to apply compression techniques to the process of discovering and learning patterns.The compression-based approach used is powerful because it can mine patterns in completely different domains. It can classify musical styles of pieces of music and identify unknown composers. It can identify the language of bodies of text. It can discover the relationships between species of life and even the origin of new unknown viruses such as SARS. Other uncharted areas are up to you to explore. In fact, this method is so general that it requires no background knowledge about any particular classification. There are no domain-specific parameters to set and only a handful of general settings. Just feed and run.CompLearn was written by Rudi Cilibrasi.**Press**
- CompLearn is described in New Scientist and Technology Research News.**About**
- CompLearn was developed at the National Research Institute CWI in Amsterdam. It was created to support the research paper Algorithmic Clustering of Music.
- **System Requirements**
- CompLearn runs under Windows and Unix. Complearn requires an installation of the Ruby scripting language. Installation instructions for Windows, Linux, and Unix here.
- Compression is achieved through the Ruby BZ2 library.
- For visualizing your graphed results, the AT&T's graphviz package is also needed.
- The toolkit requires very small amounts of disk space to install and run.
- Web design by juliob.com

You can use it too!

- CompLearn Toolkit:

  http://complearn.sourceforge.net

- "x" and "y" are literal objects (files);

  What about abstract objects like "home", "red", "Socrates", "chair", ….?

Or names for literal objects?

Non-Literal Objects

- Googling for Meaning

- Google distribution:

$$g(x) = \frac{\text{Google page count "x"}}{\text{\# pages indexed}}$$

Google Compressor

- Google code length:

$$G(x) = \log \; 1 / g(x)$$

This is the Shannon-Fano code length that has minimum expected code word length w.r.t. $g(x)$.

Hence we can view Google as a Google Compressor.

Normalized Google Distance (NGD)

- $$NGD(x,y) = \frac{G(x,y) - \min\{G(x),G(y)\}}{\max\{G(x),G(y)\}}$$
  - Same formula as NCD, using C = Google compressor
  - Use the Google counts and the CompLearn Toolkit to apply NGD.

Example

- “horse”: #hits = 46,700,000
- “rider”: #hits = 12,200,000
- “horse” “rider”: #hits = 2,630,000
- #pages indexed: 8,058,044,651

NGD(horse,rider) = 0.443

Theoretically+empirically: scale-invariant

Universality

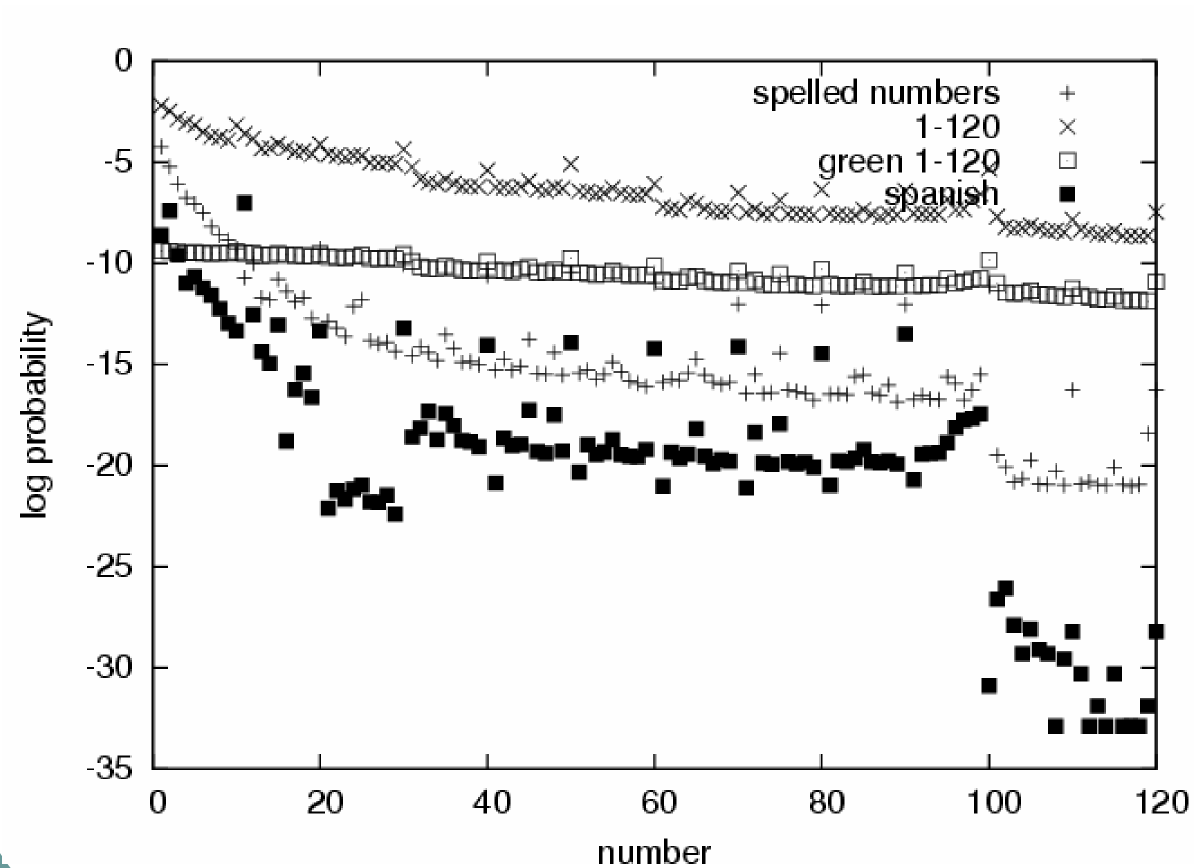
- Every web author i generates its own
  - individual Google distribution g_i
  - individual Google code length G_i
  - individual NGD denoted NGD_i

Theorem $g\_i(x) = O(g(x))$;

$G(x) = G\_i\,(x) + O(1)$;

$NGD(x,y) \leq NGD\_i\,(x,y)$, w.h.p.

Numbers versus log-probability



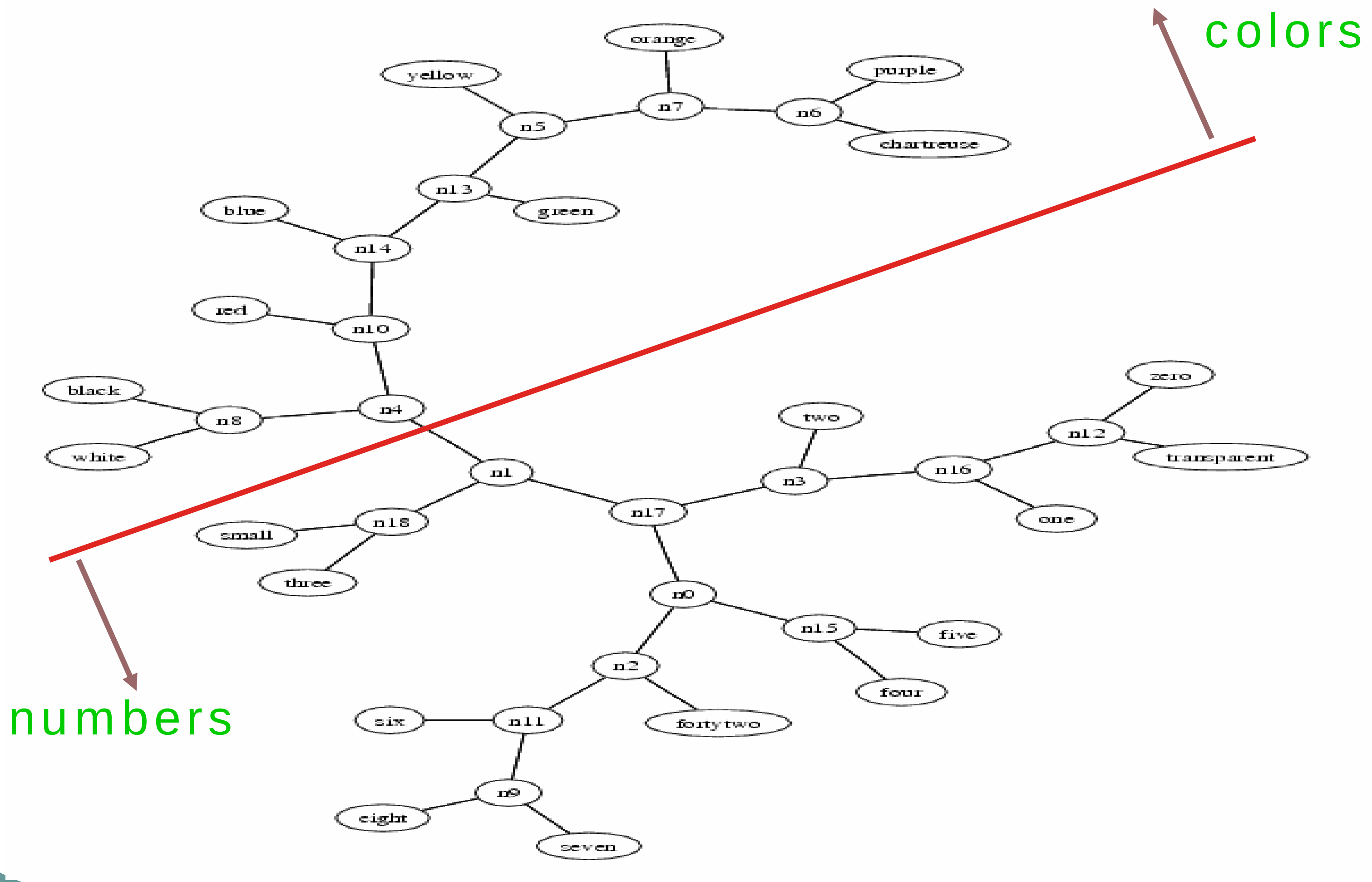
Probability according to Google.

Names in variety of languages and digits.

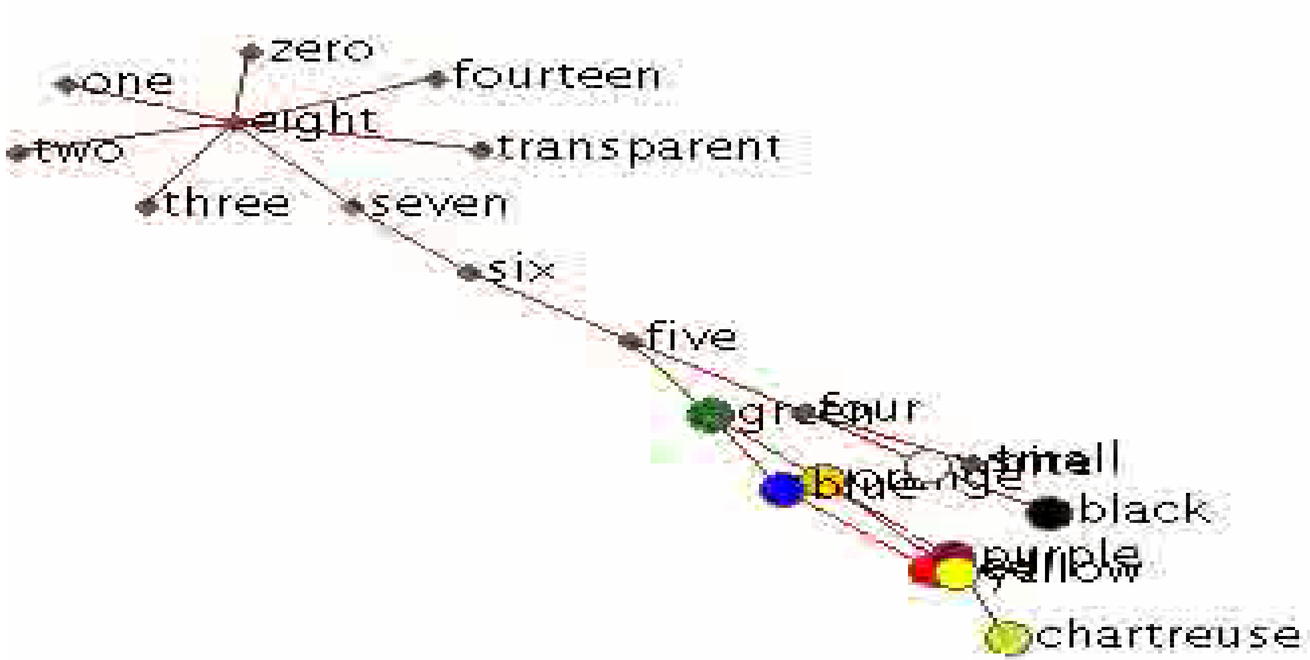
Same behavior in all formats. Google detects meaning:

All multiples of five stand out.

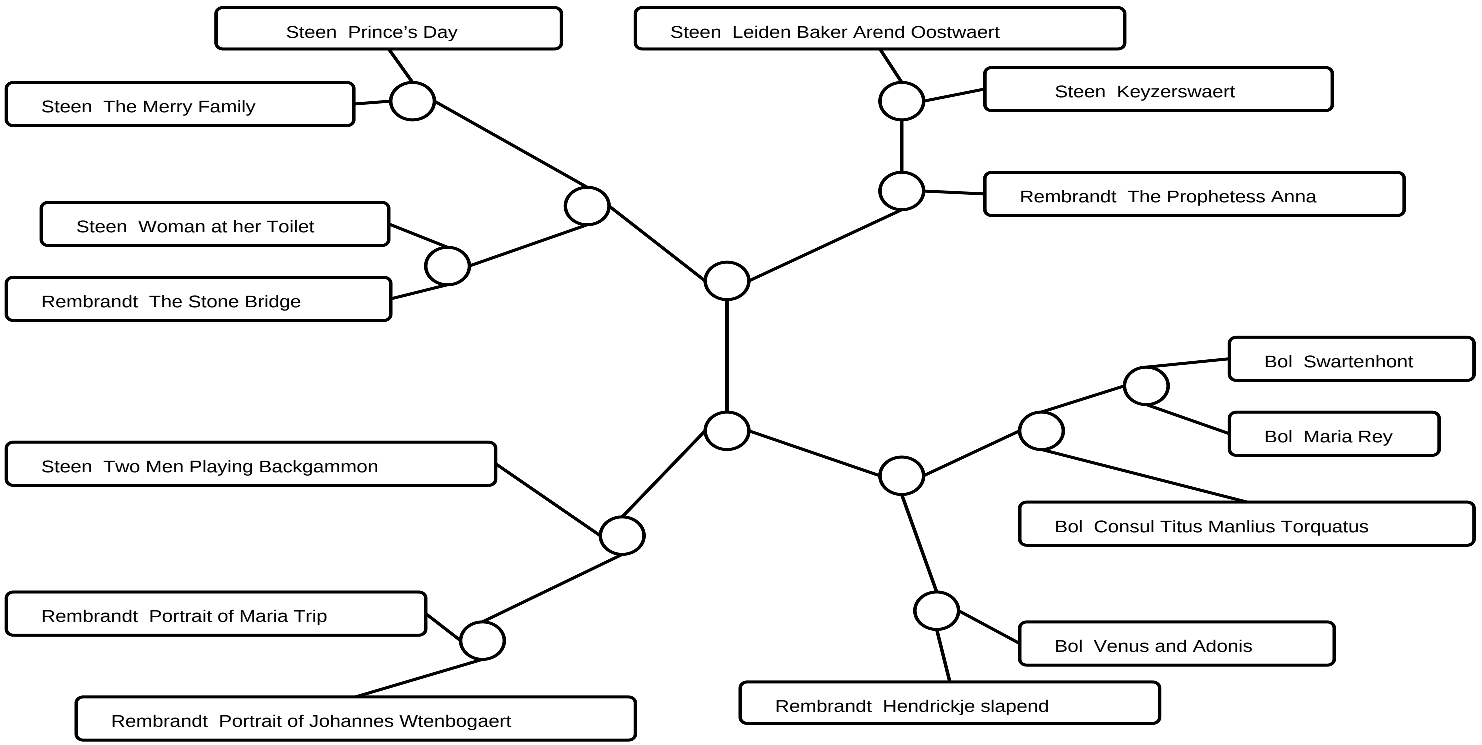
Colors and Numbers—The Names! Hierarchical Clustering



Colors vs Numbers: Minimum Spanning Tree Animation



Hierarchical Clustering of 17th Century Dutch Painters, Paintings given by name, without painter's name.



Hendrickje slapend, Portrait of Maria Trip, Portrait of Johannes Wtenbogaert, The Stone Bridge, The Prophetess Anna, Leiden Baker Arend Oostwaert, Keyzerswaert, Two Men Playing Backgammon, Woman at her Toilet, Prince's Day, The Merry Family, Maria Rey, Consul Titus Manlius Torquatus, Swartenhont, Venus and Adonis

Using NGD in SVM (Support Vector Machines) to learn concepts (binary classification)

Training Data

Positive Training	(22 cases)			
avalanche	bomb threat	broken leg	burglary	car collision
death threat	fire	flood	gas leak	heart attack
hurricane	landslide	murder	overdose	pneumonia
rape	roof collapse	sinking ship	stroke	tornado
train wreck	trapped miners			
Negative Training	(25 cases)			
arthritis	broken dishwasher	broken toe	cat in tree	contempt of court
dandruff	delayed train	dizziness	drunkenness	enumeration
flat tire	frog	headache	leaky faucet	littering
missing dog	paper cut	practical joke	rain	roof leak
sore throat	sunset	truancy	vagrancy	vulgarity
Anchors	(6 dimensions)			
crime	happy	help	safe	urgent
wash				

Testing Results

	Positive tests	Negative tests
Positive Predictions	assault, coma, electrocution, heat stroke, homicide, looting, meningitis, robbery, suicide	menopause, prank call, pregnancy, traffic jam
Negative Predictions	sprained ankle	acne, annoying sister, campfire, desk, mayday, meal
Accuracy	15/20 = 75.00%	

Example:

Emergencies

Example: Classifying Prime Numbers

Training Data

Positive Training	(21 cases)			
11	13	17	19	2
23	29	3	31	37
41	43	47	5	53
59	61	67	7	71
73				
Negative Training	(22 cases)			
10	12	14	15	16
18	20	21	22	24
25	26	27	28	30
32	33	34	4	6
8	9			
Anchors	(5 dimensions)			
composite	number	orange	prime	record

Testing Results

	Positive tests	Negative tests
Positive Predictions	101, 103, 107, 109, 79, 83, 89, 91, 97	110
Negative Predictions		36, 38, 40, 42, 44, 45, 46, 48, 49

Accuracy 18/19 = 94.74%

Example: Electrical Terms

Training Data

Positive Training	(58 cases)			
Cottrell precipitator	Van de Graaff generator	Wimshurst machine	aerial	antenna
attenuator	ballast	battery	bimetallic strip	board
brush	capacitance	capacitor	circuit	condenser
control board	control panel	distributer	electric battery	electric cell
electric circuit	electrical circuit	electrical condenser	electrical device	electrical distribut
electrical fuse	electrical relay	electrograph	electrostatic generator	electrostatic machi
filter	flasher	fuse	inductance	inductor
instrument panel	jack	light ballast	load	plug
precipitator	reactor	rectifier	relay	resistance
security	security measures	security system	solar array	solar battery
solar panel	spark arrester	spark plug	sparking plug	suppresser
transmitting aerial	transponder	zapper		
Negative Training	(55 cases)			
Andes	Burnett	Diana	DuPonts	Friesland
Gibbs	Hickman	Icarus	Lorraine	Madeira
Quakeress	Southernwood	Waltham	Washington	adventures
affecting	aggrieving	attractiveness	bearer	boll
capitals	concluding	constantly	conviction	damming
deeper	definitions	dimension	discounting	distinctness
exclamation	faking	helplessness	humidly	hurling
introduces	kappa	maims	marine	moderately
monster	parenthesis	pinches	predication	prospect
repudiate	retry	royalty	shopkeepers	soap
sob	swifter	teared	thrashes	tuples
Anchors	(6 dimensions)			
bumbled	distributor	premeditation	resistor	suppressor
swimmers				

Testing Results

	Positive tests	Negative tests
Positive Predictions	cell, male plug, panel, transducer, transformer	
Negative Predictions		Boswellizes, appointer, enforceable, greatness, planet
Accuracy	$10/10 = 100.00\%$	

Example: Religious Terms

Training Data

Positive Training	(22 cases)			
Allah	Catholic	Christian	Dalai Lama	God
Jerry Falwell	Jesus	John the Baptist	Mother Theresa	Muhammad
Saint Jude	The Pope	Zeus	bible	church
crucifix	devout	holy	prayer	rabbi
religion	sacred			
Negative Training	(23 cases)			
Abraham Lincoln	Ben Franklin	Bill Clinton	Einstein	George Washington
Jimmy Carter	John Kennedy	Michael Moore	atheist	dictionary
encyclopedia	evolution	helmet	internet	materialistic
minus	money	mouse	science	secular
seven	telephone	walking		
Anchors	(6 dimensions)			
evil	follower	history	rational	scripture
spirit				

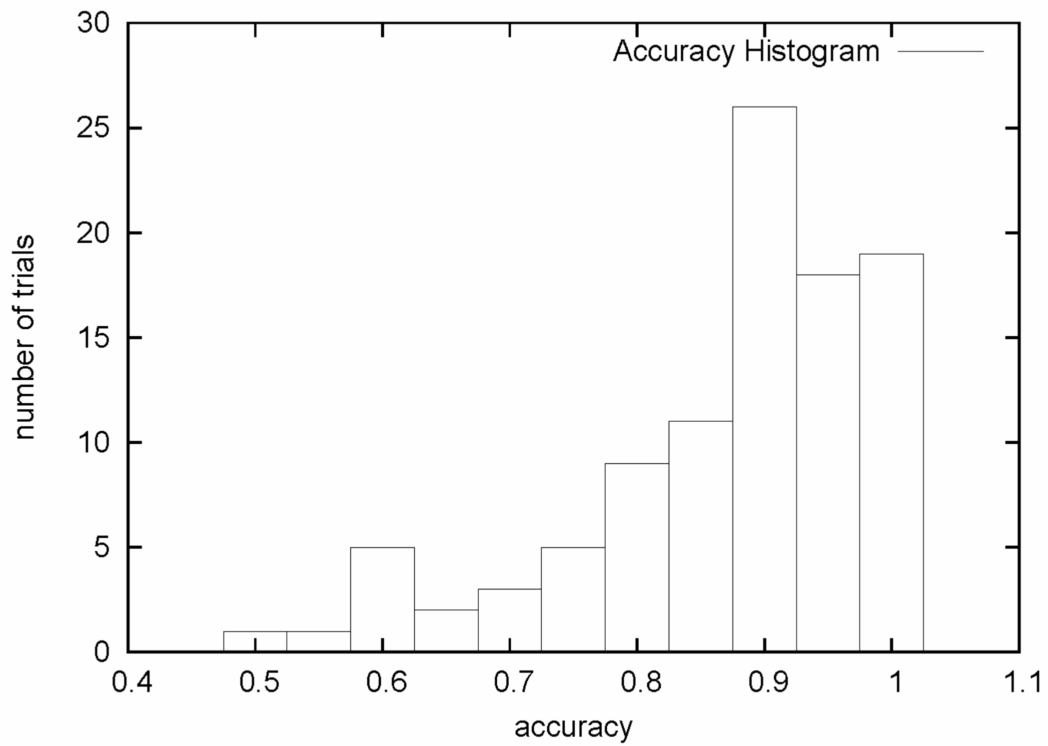
Testing Results

	Positive tests	Negative tests
Positive Predictions	altar, blessing, communion, heaven, sacrament, testament, vatican	earth, shepherd
Negative Predictions	angel	Aristotle, Bertrand Russell, Greenspan, John, Newton, Nietzsche, Plato, Socrates, air, bicycle, car, fire, five, man, monitor, water, whistle

Accuracy $24/27 = 88.89\%$

Comparison with WordNet Semantics

http://www.cogsci.princeton.edu/~wn



NGD-SVM Classifier on 100 randomly selected WordNet Categories

Randomly selected positive, negative and test sets

Histogram gives accuracy With respect to PhD experts entered knowledge in the WordNet Database

Mean Accuracy is 0.8725
Standard deviation is 0.1169

Accuracy almost always > 75%

--Automatically

Translation Using NGD

Problem:

Given starting vocabulary	
English	**Spanish**
tooth	diente
joy	alegria
tree	arbol
electricity	electricidad
table	tabla
money	dinero
sound	sonido
music	musica
Unknown-permutation vocabulary	
plant	bailar
car	hablar
dance	amigo
speak	coche
friend	planta

Translation:

Predicted (optimal) permutation

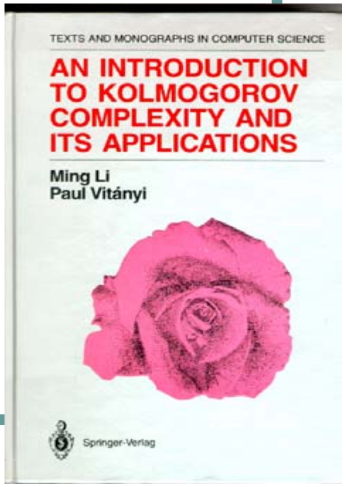
English	Spanish
plant	planta
car	coche
dance	bailar
speak	hablar
friend	amigo

Selected Bibliography

D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping, Physical Review Letters, 88:4(2002) 048702.
C.H. Bennett, P. Gacs, M. Li, P.M.B. Vitanyi, and W. Zurek. Information Distance, IEEE Transactions on Information Theory, 44:4(1998), 1407--1423.

C.H. Bennett, M. Li, B. Ma, Chain letters and evolutionary histories, Scientific American, June 2003, 76--81.
X. Chen, B. Francia, M. Li, B. McKinnon, A. Seker, Shared information and program plagiarism detection, IEEE Trans. Inform. Th., 50:7(2004), 1545--1551.
R. Cilibrasi, The CompLearn Toolkit, 2003, http://complearn.sourceforge.net/ .
R. Cilibrasi, P.M.B. Vitanyi, R. de Wolf, Algorithmic clustering of music based on string compression, Computer Music Journal, 28:4(2004), 49-67.
R. Cilibrasi, P.M.B. Vitanyi, Clustering by compression, IEEE Trans. Inform. Th., 51:4(2005)

R. Cilibrasi, P.M.B. Vitanyi, Automatic meaning discovery using Google, http://xxx.lanl.gov/abs/cs.CL/0412098 (2004)
E. Keogh, S. Lonardi, and C.A. Rtanamahatana, Toward parameter-free data mining, In: Proc. 10th ACM SIGKDD Intn'l Conf. Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22---25, 2004, 206--215.
M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny, Bioinformatics, 17:2(2001), 149--154.
M. Li and P.M.B. Vitanyi, Reversibility and adiabatic computation: trading time and space for energy, Proc. Royal Society of London, Series A, 452(1996), 769-789.
M. Li and P.M.B Vitanyi. Algorithmic Complexity, pp. 376--382 in: International Encyclopedia of the Social \& Behavioral Sciences, N.J. Smelser and P.B. Baltes, Eds., Pergamon, Oxford, 2001/2002.
M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitanyi. The similarity metric, IEEE Trans. Inform. Th., 50:12(2004), 3250- 3264.
M. Li and P.M.B. Vitanyi. An Introduction to Kolmogorov Complexity and its Applications, Springer-Verlag, New York, 2nd Edition, 1997.
A.Londei, V. Loreto, M.O. Belardinelli, Music style and authorship categorization by informative compressors, Proc. 5th Triannual Conference of the European Society for the Cognitive Sciences of Music (ESCOM), September 8-13, 2003, Hannover, Germany, pp. 200-203.
S. Wehner, Analyzing network traffic and worms using compression, Manuscript, CWI, 2004. Partially available at http://homepages.cwi.nl/~wehner/worms/



NEW SCIENTIST, 29 Jan. 2005

Technology

A search for meaning

Just google it

Using search engines to glean the meaning of words may be the vital spark leading to intelligent computers

DUNCAN GRAHAM-ROWE

COMPUTERS can learn the meaning of words simply by plugging into Google. The finding could bring forward the day that true artificial intelligence is developed.

Trying to get a computer to work out what words mean – distinguish between "rider" and "horse" say, and work out how they relate to each other – is a long-standing problem in artificial intelligence research.

One of the difficulties has been working out how to represent knowledge in ways that allow computers to use it. But suddenly that is not a problem any more, thanks to the massive body of text that is available, ready indexed, on search engines like Google (which has more than 8 billion pages indexed).

The meaning of a word can usually be gleaned from the words used around it. Take the word "rider". Its meaning can be deduced from the fact that it is often found close to words like "horse" and "saddle". Rival attempts to deduce meaning by relating hundreds of thousands of words to each other require the creation of vast, elaborate databases that are taking an enormous amount of work to construct.

But Paul Vitanyi and Rudi Cilibrasi of the National Institute for Mathematics and Computer Science in Amsterdam, the Netherlands, realised that a Google search can be used to measure how closely two words relate to each other. For instance, imagine a computer needs to understand what a hat is.

To do this, it needs to build a word tree – a database of how words relate to each other. It might start off with any two words to see how they relate to each other. For example, if it googles "hat" and "head" together it gets nearly 9 million hits, compared to, say, fewer than half a million hits for "hat" and "banana". Clearly "hat" and "head" are more closely related than "hat" and "banana".

To gauge just how closely, Vitanyi and Cilibrasi have developed a statistical indicator based on these hit counts that gives a measure of a logical distance separating a pair of words. They call this the normalised Google distance, or NGD. The lower the NGD, the more closely the words are related.

By repeating this process for lots of pairs of words, it is possible to build a map of their distances,

> "The web might make all the difference to whether we make an artificial intelligence or not"

indicating how closely related the meanings of the words are. From this a computer can infer meaning, says Vitanyi. "This is automatic meaning extraction. It could well be the way to make a computer understand things and act semi-intelligently," he says.

The technique has managed to distinguish between colours, numbers, different religions and Dutch painters based on the number of hits they return, the researchers report in an online preprint (www.arxiv.org/abs/cs.CL/0412098).

The pair's results do not surprise Michael Witbrock of the Cyc project in Austin, Texas, a 20-year effort to create an encyclopaedic knowledge base for use by a future artificial intelligence. Cyc represents a vast quantity of fundamental human knowledge, including word meanings, facts and rules of thumb. Witbrock believes the web will ultimately make it possible for computers to acquire a very detailed knowledge base. Indeed, Cyc has already started to draw upon the web for its knowledge. "The web might make all the difference in whether we make an artificial intelligence or not," says Witbrock. ●