



## Raul Valdes-Perez, co-founder & President

- 1991 Ph.D. in Computer Science (Carnegie Mellon)
- CMU CS Faculty member since then, currently on leave
- Research expertise in AI and Knowledge Discovery

## > Vivisimo, Inc.

- Enterprise software company founded June 2000 in Pittsburgh
- Award-winning web-search site Vivisimo.com
  - Best meta-search site last 2 years in row (Search Engine Watch)
  - Site is just a demonstration of our clustering & EAI products
- Main Funding: National Science Foundation

## **Problem: Information Overload & Overlook**

## > Information Overload and Information Overlook

- Look for information  $\rightarrow$  get too much back
- Most people handle overload by overlooking most information

Vivísimo



## > Overlooking information has a business/user cost

- Corporate employees fail to find needed information
- Customers don't solve their problems by themselves, online
- Publishers lose potential readership
- Web search providers lose potential click-throughs
- Users miss out on unexpected discoveries or opportunities



I'm Feeling Lucky

# How to Alleviate Information Overlook?

Vivísimo

## > Decrease the available information

- Purge the obsolete
- Censor the worthless

## Make people ...

- Smarter
- Work harder
- More efficient! (but how?)

## > Provide organized information!

## How to Provide Organized Information?



#### Manually tag (classify, index) all content?

• "We have no process for consistently tagging our content. We have 50 different business units. People in one unit do a great job, but others do not use tags at all." <u>Forrester Report</u>

Vivísimo

- Forrester says \$4 per page to make a controlled vocabulary
- \$50 per document to manually tag (large pharma)
- Tags tend to be broad and bland (one size fits all)
- → Tagging is costly and leads to mediocre results

## > 1. Spatial approaches (also temporal)

- Assign documents to spatial or temporal coordinates
- Intuition: people deal well with space/time so embed documents in space-time
- Simple space: 2D maps (search lists are like 1D space)
- Prominent example: <u>Pacific Northwest Labs</u>
- Many commercial variants

## > 2. Clustering & categorization

- Assign documents to discrete, named groups
- More than one group is OK & can be hierarchical
- Intuition: people use groups/hierarchies all the time
- Prominent examples: <u>NorthernLight</u>, <u>Vivisimo</u>

# **1. Spatial Approaches**



#### > Use vector-space representation of documents

- The vector dimensions are the document words & phrases
- Values are #times that a word or phrase appears
- Very high-dimensional space
- Each document is a point in this space

#### > Then, select 2-3 axes that best scatter the points

- E.g., principal components analysis
- Embed the documents in this 2-3D space
- Show the map

#### Pros

• Can handle a very large number of documents

#### Cons

- Textually characterizing the spatial areas is difficult
- Drill-down is clumsy

# 2. Categorization and Clustering

#### Categorization

- Places documents in pre-defined categories
- NorthernLight
  - 1. Librarians build/maintain controlled vocabulary of categories
  - 2. Software is designed/trained to assign documents to categories

Vivísimo

- 3. When querying database, resolve returned document sets into the most frequent categories
- Yahoo directories & <u>Librarian's Index to the Internet</u>
  - Like above, but step (2) is done manually

#### Clustering

- Groups documents and makes spontaneous descriptions for them
- Many years of work in academia & industry
- Example: <u>IBM's Intelligent Miner</u>
- <u>Vivisimo Clustering Engine</u>

# **Pros/Cons of Using Pre-Defined Categories**

## Pros

- Guarantees human-like categories
- Algorithm is simple, once the categories have been created and assigned

Vivísimo

#### Cons

- Expensive to develop/maintain (staff time, etc.)
- Scales poorly across human languages (German, Chinese)
- Typically must pre-process all the records
  - Some resources are accessed only by searching them
  - Can be logistically complicated
- Automatic classification into categories is error-prone
- Reduced chance of surprise/discovery
- Fixed vocabularies tend to be broad and bland
  - Queries on "Three Rivers Stadium" or "Kingdome"
  - Fixed categories might place in "Construction Industry"
  - Vivisimo says "Implosion" or "Memories of Three Rivers Stadium"

# **Document Clustering is Hard to Do Well**

#### > Cluster analysis in general is not so hard

• e.g., Warehouse location problem

#### > Document clustering is not so hard when ...

• People don't need to understand the clusters

#### Clustering for post-retrieval browsing is hard

• People need to understand the clusters at a glance

#### Most doc-clustering algorithms optimize the wrong things

• Mathematical abstractions are distant from the main quality factors

Vivísimo

#### > Quality of description is paramount

- Concise, accurate, natural, hierarchical, and others
- Hard to formulate mathematically, so develop heuristic algorithm!
- Needs to use statistics, linguistics, and subject matter knowledge

# Modeling the Value of Topic Clusters



- Lots of wasted effort if information is disorganized
- View few results before exhausting your time allotment

Vivísimo

#### Modeling Assumptions

- User spends 12 min before giving up or moving on
- Eye skips over search results or folders sequentially
- Equal, independent probability that any given result solves the problem
- Employee costs \$60 per hour
- 1,000 employees
- 2 searches per employee/day
- 10 minutes to solve problem elsewhere when search fails
- **Folders let you see 11 docs in detail vs. 6 for ranked lists**
- Conclusion: savings of \$1M+ per year (white paper)



- Users click on 2.8 folders, on average (publisher customer)
- > Increased click-throughs of 30% to 200%
  - 30% increase at vivisimo.com by adding up to 28 sponsored results at the bottom of 200+ total search results
  - Essentially all of these are accessed via the folders
  - Larger increases when the BEFORE is unclustered ranked lists

#### Fortune 1000 customer experimental tests

- Combination of meta-search and clustering
- Minutes-to-solve: 1.9 versus 3.7
- #Searches-to-solve: 1.4 versus 2.3
- Rating: 1.6 versus 2.4 (1=very easy, 5=impossible)

## Hybrid Approach May Offer Best of Both Worlds

#### > Automatic clustering on several fields

- Document title
- Summary (aka abstract, snippet, excerpt)
- Humanly assigned, pre-defined categories

#### > Doesn't matter if some fields (e.g. categories) are missing

Vivísimo

• Just cluster on what's available

#### > Dynamic (not fixed) categories will prevail ...

- When they offer better folder descriptions than pre-defined categories
- At deeper points in the folders hierarchy, where pre-defined categories are not detailed enough

#### Example: <u>clustering PubMed search results</u>

• Clusters on titles, abstracts, and Mesh Headings

# How the World is Now



Google	Advanced Search Portices Concern	ces Language Tec	is SeechTas Geogle Search						
Web Integer Groups Searched the web for gamper.	Harrison's On	line		Min a					
Category Health > Conditions	Braf Cartara	Access@edicionHer	Merck.com Search	Results					
Nove: Breath cancer research Star Tribute - 3 hours ag	Pull Contents Update: Clinical Trials	Search Re	Start new search Search these results			-			
Natural Insecticide Identif have ago Doast concert land and	Reviews & Editorials Res Tabla	Search for can	Search:	Pub Med	of Medicine	(L.54			
5 heuro age New Try Geogle News 1	Related Science Forum	Would you like	passe.	ulardih Przah Gerona Strutte	Search RESOLTS	You reached on a	North Tel	al seconds matched = 50	87 of 30136
American Cancer Society Ho	Tables and Figures (31)		Results for: cancer	≝ <sub>for</sub> oancer		Peccilit Matched	DEPENDING NUMBER	Cancer united life	ander 1800
Buttan, Making Strides Again	Refine Search	Chapters /I	The query is assumed to be in English.	Lines Preventides Pie	To all administration sections.	2124	Rearch willin Pass results	Cancer Corry In	100
Description: Dedicated to helping		Carapter ay 1	101 months for and constant has enhanced	Display Summary * Sort	Page reviewed journal	Recent Matheet	@ Bearsh by: 17LE OBLI	C. Bearsh ky: 10	LE & TEXT
services, early detection, Category: Health × Canditions at	R double With	Chapter EO: Pr	the research round, seems by remained	- 20 - Dame 1-20 of 3444115	Dustations	Force Matcheol 200	Non Search		00
www.catcar.org/-31k - Oct. 16,	C teach trug List	Otto W. Bnawler Part & Oncolo Score: 1.00	THE MERCK MANUAL, Sec. 11, Ch. 164, Princi Additive or ablative endocrine therapy can influe	Star 10	eTechnoles and readical images	Precents Matchiet 2067	gt: Search by: 111LE ONL1	C Search by: TH	LE & TEXT
Cancer gov text only skip navigation. National	Can		t is only	These E. Damsti CA. Lakeh H. Conserve LL. South	Experimental links to Clinical I	Queries on PubMed.			
cancer information clinical trials		Chapter BO: Pt		Am   Ind Med. 2002 Nov (42(5):410-420.	Evidence-Based (Direct Index) and effect method. West the has	The links on the Mds section of both St.	reme wen Beitigene banefferenerf, i mit Bellere bes Derberent eine 1982 (b	a they have been provide	ally appraised using a initial firms. Filmula
Calegory: Health > Canditions at		Part & Oncolo	GU tumots occur at any age and in both serves.	PRES 120204 POINT - or repairs by	Obesity - Problems and inte	mentions.			2002
www.nci.nit.gov/~25k - Oct. 16,		Score: 1.00	in the male and 4% in	T 2-Roberto ME, Offic K.	Resultati Council on Technology de Ostatikationeses: De for planet e	esses worth			
<b>Oncolumis</b>		Chapter 97: G	age 20.44	Considerations in genetic counseling for	Gignpostfilleer	A DIVINE CANCEL			2002
Concer News, Het Spot. Tuer 13 deaths related to concer dep Description: Description and an area		Part & Oncolo Score: 1.00	THE MERICK INVALIAL, Sec. 11, Ch. 144, Princi A common misconception is that some cancer which each in textual	PME: 1282194 [PubMed - as supplied by:	Tipping the belience?: lower r SAEA/S	tsk of childhood branchi	al authoria is a further reasons to	o breast-feed	2002
Coloppy, Health. Support Drs.			pierent can be treated pize 24.0K	E 3-frenzia	Fony Fate Feats Real Defeat of Physics and Londo	IN .			2002
www.oncolink.com/ - 24k - Oct.		Chapter 89: 84	THE BERCK HANDAL, Sec. 11, Ch. 144, Princh	Rolistherapy for breast cancer in BRC	Key recommendations and a Read College of Physicsen Londo	neniew of management	of theroid cancer		2002
Cancer Care Cancer Care provides the profet		Part & Oncolo Score: 1.00	The limited success of treatment with surgery o with radiation can	diletations. Swain Radiat Oncol. 2003 Oct.) 3(4):153-364.	Early Goal-Directed Therapy Point Dirithmore Based Journal	is the Treatment of Seve Chile	ee Sepsis and Septia Shock		2002
counceling, education, referral or			aize 23.8K	FREX 120200 (Poisses - ar repose by)	Camplications of Femeral and	el Sabolavian Venous Ca	theterization in Critically III Pa	tionts: A Randomized	2000
Category. Health > Canditions.at		Editorial Rolati	THE MERCK MANUAL, Sec. 11, Ch. 144, Princi	E 4-Status CL	PastaC DM Rvidence Based Journal	Chuk			
www.cancercare.org/ - 22k - Oct		Postect 09-01 Score: 1.00	The ideal chemotherapeutic any would target a normal cells alse 3d.2K	Molecular targets as therapeutic strateg Semin Reduit Oncel. 2003 Oct.12(4):345-354. DMTC 12(2020) DeatMod	Noninvasivo ventilation in ime respiratory follow Petro Del Evidence Parel Journal	nunesuppressed patient	s with pulmonary infitrates, for	er; and acute	2002
cancomevs.com is a site dedica		EDITORIAL: BA	THE MERICK MANUAL Sec. 11 (b. 14) Provi	y many an appare of t	Advanced Celerectal Cancer	- innotacian, exalipitatin-	& ratificated		2002
www.cancernows.cem/~4k + Ea		Editorial Relati Dan L. Longo	Radiatherapy can be delivered by various swithe	E Stillefly BD.	Eneral cancer - Institutureab				2002
		Postect 03-12 Score: 0.99	which largely delivers aloe 25.2%	Molecular and genetic markers in the in Semin Barliat Oncol. 2003 Oct.1204(338-340.	National Institute for Clinical Excel	New Coastern			
			THE BETTEK BANDAL Harry Edition Sec. 22.	PMID: 12082194 [PubMed- as supplied by:	Haltonal Inditute for Clinical Excel	letan			2000
		Editorial Relati Inflammatory	Increasing age is an important tisk factor. About prestant offer age 75. Previous	C 6 Crim II. Data AA. Cartofindi M. Arm B. H	Can initial prostate specific a revely diagnosed prostate ca 1993 Boseneris Bushatter Databas	istigen determinations e ecinoma? A multicenter #	insingle the need for bone sos retrospective study in Japan	ns in patients with	3005
			1979 20.3K	Senin Radiat Oncol. 2002 Oct;12(4):318-328.	Pill reduces risk of overlas of Hilling the Hamilton	ancer			2002
			UNE MERICK INVOLUL-Home Edition, Sec. 15.	CHEFY TENTON LARGER - IN MIDDING (Å.)	Nan-Hodgkin's lymphoma de Hilling the Headlines	ug break Straugh			2002
				7:50 an DA. Melacular reactors in the documents and	Routine teaching of breast section, the teaching	elf examination does not	save lives		2002
				Swean Radiat Owent, 2003 Over 3 2(4) 205-318. PMED: 12382109 [PubMed - as regulard by:	Risks of hormone replacements Histographic Heathers	et therapy			2002
					Aspirin will put terresifies risk Hilling the Headlose	ku"			2002
					Telestane is the treatment of	advanced liveast cance	63		2002

# What Vivisimo Makes Possible



<b>∀</b> ▼ Vivísimo	cancer	Search					
Bio MetaCluster	☑ PubMed @NIH ☑ I ☑ TRIPDatabase	MerckManual 🗹 Google 🗹 Harrison					
Clustered Results		Category cancer > Colorectal cancer > Advanced colorectal cancer contains 5					
<u>cancer (</u> 571)		documents.					
⊕ ► <u>Breast cancer</u> (129)							
⊖ <mark>∀ Colorectal cancer</mark> (62)		1. Combination chemotherapy for advanced colorectal cancer. [New Window] [Full					
Early Detection (19)		Window] [Preview]					
Colorectal Cancer Screening	9)	British Journal of Cancer (2003) 88, 1152-1153. doi:10.1038/sj.bjc.6600848					
Predictors, Colorectal adenon	ias (5)	Authors: Mason M - Johnson P - Rudd R -					
Advanced colorectal cancer (5	)	HRL: www.nchi.nlm.nih.gov:80/entrez/guery_fogi?omd=Retrieve&db=Pu					
Chemotherapy For Advance	, d Colorectal Cancer (2)	Source: PubMedXML 41th					
⊕ → Guidance, Irinotecan, Oxali	latin & Raltitrexed (2)						
• T lymphocytes isolated from natie	ts with advanced colorectal c	2. Comparison of intermittent and continuous palliative chemotherapy for					
⊕ ► Risk of Colorectal Cancer (5)		advanced colorectal cancer: a multicentre randomised trial [New Window] [Full					
Screening for colorectal cancel	ar (3)	Window] [Preview]					
(h) Liver metastases (2)	<u>~</u> (*)	URL: www.tripdatabase.com/redirect.cfm?id=181233&criteria=cancer					
(2)		Source: IRIPDatabase 91th					
Colonascony, Neonlasia (2)		3 Full guidance on irinotecan, ovaligiatin & raltitraved for Advanced Colorectal					
S Nice Denky (2)		Cancer INew Window! IFull Window! IPreview!					
• Fince, Reply (2)		URL: www.tripdatabase.com/redirect.cfm?id=181100&criteria=cancer					
Other Topics (13)		Source: TRIPDatabase 7th					
Eung cancer (34)							
Gastrointestinai Tract (31)		4. T lymphocytes isolated from patients with advanced colorectal cancer are					
Prostate cancer (30)		suitable for gene immunotherapy approaches. [New Window] [Full Window] [Preview]					
⊕ ► <u>Ovarian cancer</u> (19)		Despite improvements in treatment, the 5-year survival for metastatic colorectal cance					
• Cervical cancer (19)		remains poor. Novel approaches such as gene immunotherapy are being investigated					
Principles Of Cancer Therapy (1)	3)	nuprove treatment. Retrovital gene transfer methods have been shown to transduce primary human T lymphocytes effectively resulting in the expression of therapoutio					
⊕ ► <u>Cancer Institute</u> (16)		denes. However, a number of defects have been identified in T lymphocytes isolated fr					
⊕ ► <u>Mortality</u> (16)		patients bearing tumour, which may have critical implications for the development of					
* More		gene-targeted T cells as an anticancer therapy. To address this issue, primary T					
		Lymphocytoc wore isolated from nationts with advanced colorectal concert and tested					

# Multiple Sources = Multiple Searches and Multiple Lists



## Without Vivísimo

I need to find.....

#### Problems

- 5 separate searches
- 5 individual information lists
- Relevant info stays hidden



# A New Standard, A Better Way



#### Faster, more efficient knowledge discovery.



# Value Proposition of Clustering



- Easy access to useful but low-ranked results
- Learn at a glance the types of available information

Vivísimo

See results in context of similar results

## Corporate

- More efficient employees/customer support operations
- Customers become more engaged with your content

## **IT Dept**

- > Quick installation on any search engine/database
- > Overlays search; is non-invasive; no maintenance
- No need to train users





- > Information Overlook imposes high opportunity costs
- > Alleviate by creating and showing organized info
- > Clustering into thematic folders is a natural approach
  - Has proven hard to do well over many years of research
  - But is now largely a solved problem
- Document clustering is Vivisimo's founding technology
  - "We help enterprises organize information from anywhere, any time, in any language, without the endless cost and complexity of building information taxonomies."