Une introduction au Text Mining et à la sémantique

Seth Grimes

--

Président, Alta Plana

Alta Plana

New York Times

9 octobre, 1958

NEHRU PROMISES BORDER DEFENSE

Warns Peiping India Would Resist Further Advances Into Her Territory

By PAUL GRIMES
Special to The New York Times.

NEW DELHI, India, Oct. 8—Prime Minister Jawaharlal Nehru warned today that any further aggression by Communist China against India "will certainly be fully resisted."

"I hope it will not be necessary," he said at a crowded news conference. But he warned against "any kind of advance" by the Chinese from positions they now hold inside India's northern borders.

The Prime Minister asserted, "We do not intend to start military operations against any of these places at this stage, when we are dealing with this matter on a political level."

He said the Indian Government did not function "in an excited way with a club in hand" but instead sought to act "with determination, not with anger."

Mr. Nehru recalled that he had sent a letter to Premier Chou En-lai of Communist China Sept. 26, appealing to the Chinese to withdraw their forces so that the border disputes could be "amicably and peacefully settled."

The Prime Minister said he welcomed the "friendly message" sent by Mr. Chou in response to India's greetings on the tenth anniversary of the Communist regime Oct. 1. In this message, made public here last night, Mr. Chou stressed the "age-old friendship" between China and India and described the present difficulties as "merely an episode."

"The tone is friendly and, therefore, an improvement," Mr. Nehru said. "How far that represents any basic change in the situation, I cannot say. I hope it does."

The Prime Minister, dressed in white homespun with his usual red rose in a buttonhole, answered questions for an hour and a quarter.

Prime Minister Nehru belittled a report that Premier Khrushchev, when visiting Los Angeles, told of the Soviet interception of a highly confidential message from President Eisenhower to Mr. Nehru. The Prime Minister said the only messages he had received from the United States President in the last two or three months were "friendly and formal letters—nothing secret."

Mr. Nehru also expressed the following views:

¶That he is opposed to disbanding the armistice control commissions in Laos, Cambodia and Vietnam as long as it is possible to carry out the Geneva agreements of 1954 that ended the Indochinese war.

¶That President Charles de Gaulle's latest proposal to end the Algerian strife was "certainly a marked advance" from previous French suggestions, particularly since it "acknowledged the right of self-determination, which was the basic thing."

¶That he disapproved thoroughly of a recent "extraordinary" resolution of the Indian Communist party that would have India weaken considerably her stand on the border issue.

¶That the Soviet Union should be "warmly congratulated" for its rocket around the moon.

Mr. Nehru, who will be 70 years old Nov. 14, was asked whether "you feel a sense of fulfillment in your life or a sense of frustration."

He replied:

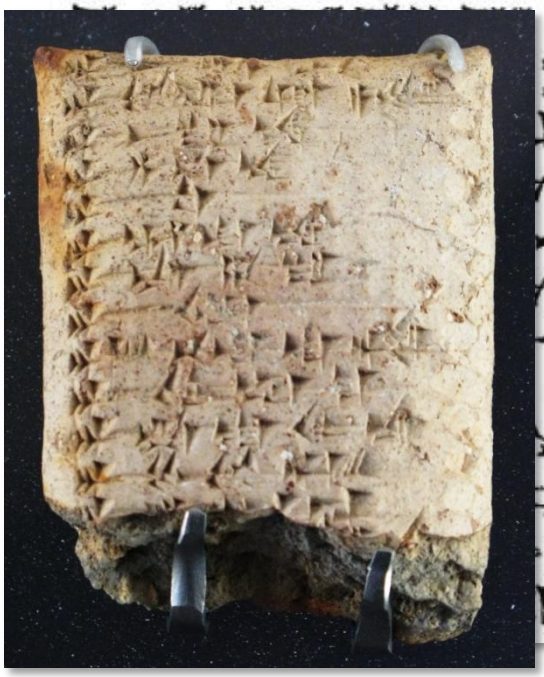
"I have absolutely no sense of frustration in my life. I hope my face shows that. If the question is, have I achieved all I want to, my answer is no. But partial achievement comes from time to time."

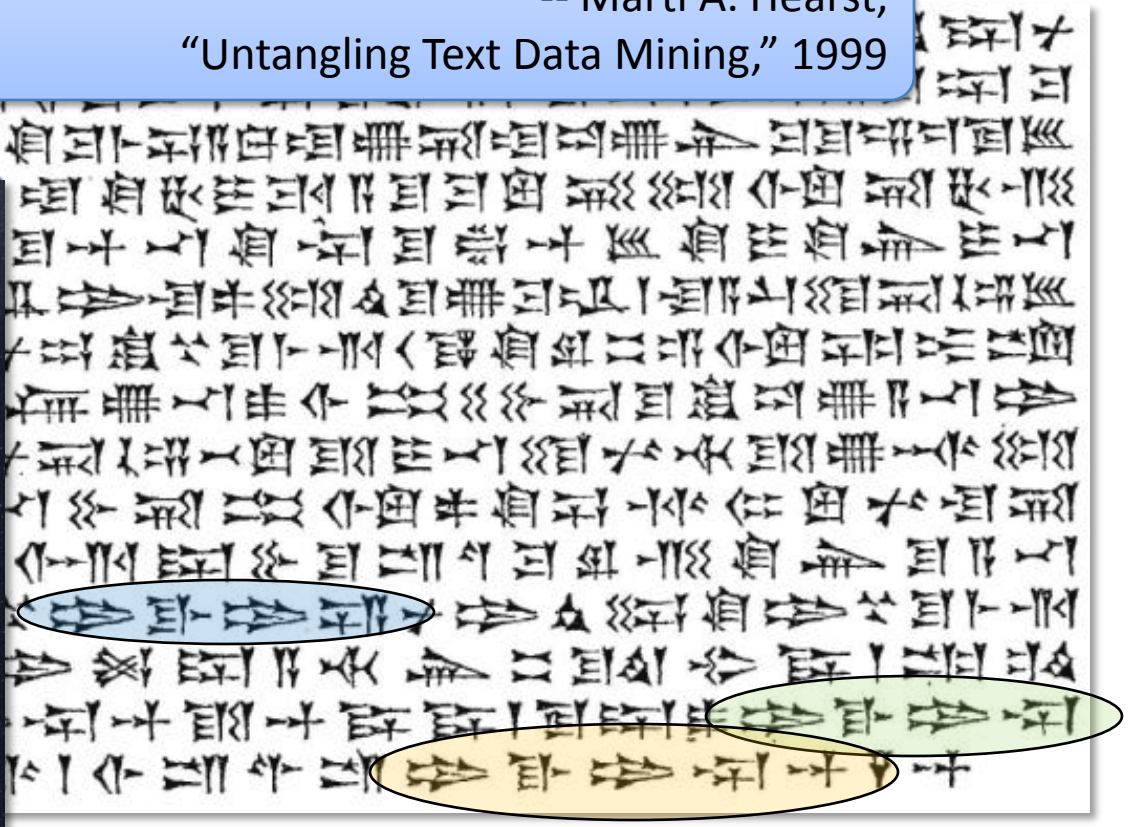
The New York Times
Published: October 9, 1959

Du texte à l'information

«Le texte exprime une gamme vaste et riche d'information, mais encode cette information dans une forme qui est difficile à déchiffrer automatiquement.»

-- Marti A. Hearst,
"Untangling Text Data Mining," 1999





Alta Plana

Input et traitement des documents

Gestion des connaissances

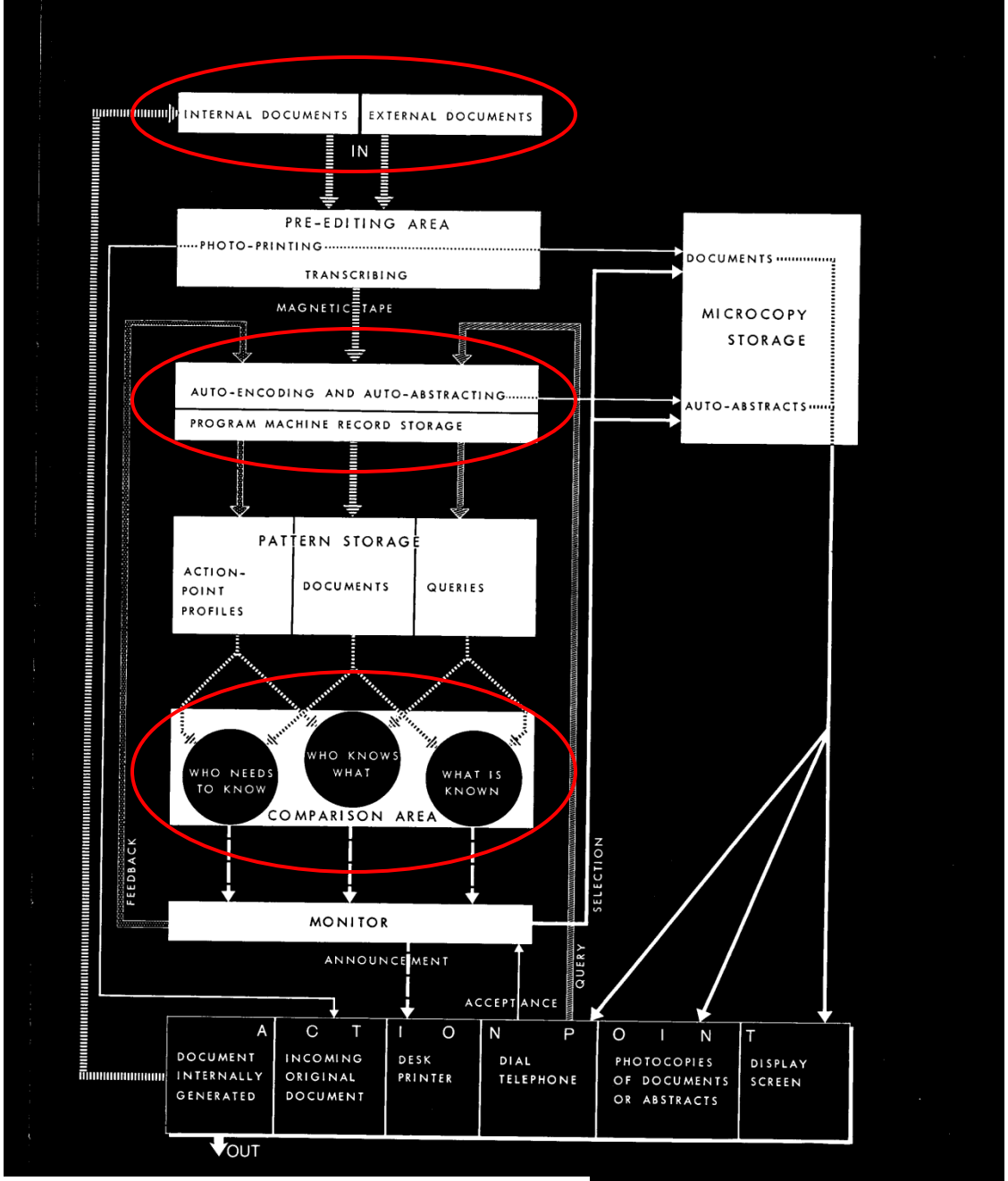


Figure 1 A Business Intelligence System

Extraction d'information

Hans Peter Luhn, "*A Business Intelligence System,*" IBM Journal, October 1958

Analyse statistique du contenu

«L'information statistique obtenue par la fréquence et la distribution des mots est utilisée par la machine afin de calculer une mesure relative de leur importance.»

Significant words in descending order of frequency (common words omitted).

46	*nerve*	12	*body*	6	*disturbance*	4	*accumulate*
40	*chemical*	12	*effects*	6	*related*	4	*balance*
28	*system*	12	*electrical*	5	*control*	4	*block*
22	*communication*	12	*mental*	5	*diagram*	4	*disorders*
19	*adrenalin*	12	*messengers*	5	*fibers*	4	*end*
18	*cell*	10	*signals*	5	*gland*	4	*excitation*
18	*synapse*	10	*stimulation*	5	*mechanisms*	4	*health*
16	*impulses*	8	*action*	5	*mediators*	4	*human*
16	*inhibition*	8	*ganglion*	5	*organism*	4	*outgoing*
15	*brain*	7	*animal*	5	*produce*	4	*reaching*
15	*transmission*	7	*blood*	5	*regulate*	4	*recording*
13	*acetylcholine*	7	*drugs*	5	*serotonin*	4	*release*
13	*experiment*	7	*normal*			4	*supply*
13	*substances*					4	*tranquilizing*

Total word occurrences in the document:	2326
Different words in document:	
Total of different words	741
Less different common words	170
Different non-common words	571
Ratio of all word occurrences to different non-common words	~4:1
Non-common words having a frequency of occurrence of 5 and over:	
Total occurrences	478
Different words	39

163

Hans Peter Luhn,
"The Automatic Creation of Literature Abstracts,"
IBM Journal, April 1958

Limites de l'analyse statistique

«Cette argumentation assez simple sur la 'signification' ignore les aspects linguistiques tels que la grammaire et la syntaxe... Aucune attention n'est accordée aux rapports logiques et sémantiques établis par l'auteur.»

-- Hans Peter Luhn, 1958

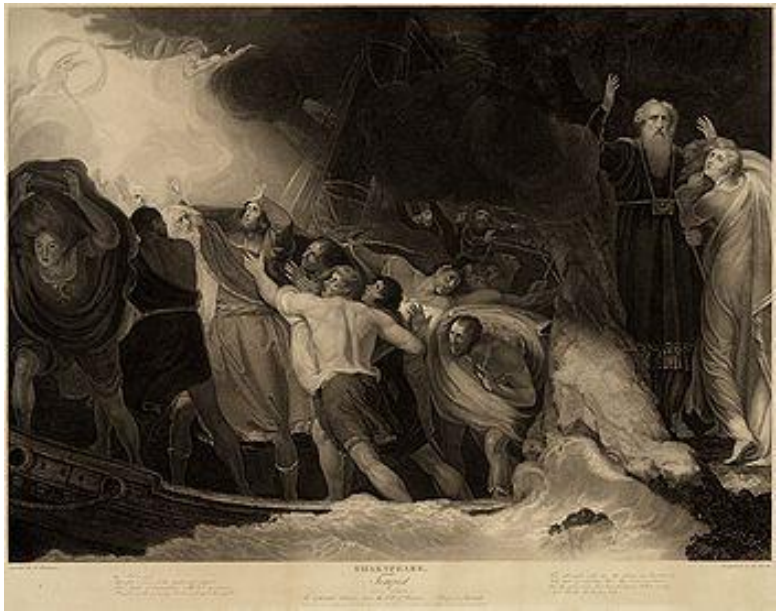
Inconnu = nouveauté

Miranda:

> O, merveille! Combien de belles créatures vois-je ici réunies! Que l'humanité est admirable! O splendide Nouveau Monde, Qui compte de pareils habitants !

Prospero:

C’est nouveau pour toi.



Le naufrage dans **La Tempête**, Acte I, Scène 1, dans une gravure de 1797 basée sur une peinture de George Romney.

Liens sémantiques

New York Times,
8 septembre, 1957

Anaphore / coréférence: “They”

SCIENCE IN REVIEW

Chemistry Is Employed in a Search for New Methods to Conquer Mental Illness

By ROBERT K. PLUMB

By coincidence this week-end in New York City marks the end of the annual meeting of the American Psychological Association and the begining of the annual meeting of the American Chemical Society.

Psychologists and chemists have never had so much in common as they now have in new studies of the chemical basis for human behavior. Exciting new finds in this field were also discussed last week in Iowa City, Iowa, at the annual meeting of the American Physiological Society and at Zurich, Switzerland, at the Second International Congress for Psychiatry.

Two major recent developments have called the attention of chemists, physiologists, physicists and other scientists to mental diseases: It has been found that extremely minute quantities of chemicals can induce hallucinations and bizarre psychic disturbances in normal people, and mood-altering drugs (tranquilizers, for instance) have made long-institutionalized people amenable to therapy.

Money to finance research on the physical factors in mental illness is being made available. Progress has been achieved toward the understanding of the chemistry of the brain. New goals are in sight.

At the psychiatrists meeting in Zurich last week, four New York City physicians urged their colleagues to broaden their concept of "mental disease," and to probe more deeply into the chemistry and metabolism of the human body for answers to mental disorders and their prevention.

Blood May Tell

Dr. Felix Marti-Ibanez and three brothers, Dr. Mortimer D. Sackler, Dr. Raymond R. Sackler and Dr. Arthur M. Sackler cited evidence that the blood chemistry of victims of schizophrenia is different from that of normal people. Perhaps multiple biological factors are responsible for this chemical change, they suggested.

Mental disease is a "developmental process" and long duration of a disorder may result in "permanent alteration of anatomy and physiology," they said. They urged that trials of new drugs which affect the brain should be concentrated on complex studies of the mechanism of action of the drugs. The variety of substances capable of producing profound mental effects is a new armory of weapons for use in investigating biological mechanisms underlying mental disease, they said.

The sources of behavioral disturbance are many and they may come from external as well as internal forces, the four reported. This concept has already proven practical, for instances, when it enabled psychiatrists to predict that the administration of ACTH and cortisone could produce psychosis.

"It led some years ago to the development of a blood test which was 80 per cent accurate in the identification of schizophrenic patients," they said. "It permitted us on physiologic grounds to deny that the psychoneuroses and the psychoses were lesser and greater degrees of the same disease process, and, in fact, to affirm that they represented opposite and even mutually exclusive directions of physiologic disturbances," they said.

Chemicals now available should be used not only to bring relief to the mentally sick but also to uncover the biological mechanisms of the disease processes themselves. "Only then will the metabolic era mature and bring to fruition man's long hoped for salvation from the ravages of mental disease," they reported.

Chemistry of the Brain

At the psychologist's meeting here, a technique for tracing electrical activity in specific portions of the animal brain was described by rese... of Cali... reporte... in cat ... electric... anima... which ... In this ... reporte... sequen... its vari... ... S... the brain may be located. Furthermore, the electrical pathways so traced out can be blocked temporarily by the use of chemicals. This poses new possibilities for studying brain c... and sic... the Ca... sized.

The ... try ha... peutic ... courage ... for me... that kn... ary fiel... last we... Washin...

This ... the Na... Health ... formati... Literati... find an... bers ca... technic... People ... vited to... or othe... letters ... have in... Informa... Silver Spring, Md.

Analyse de la tonalité

"Kind" = genre, variété, pas une indication de sentiment.

Référence externe

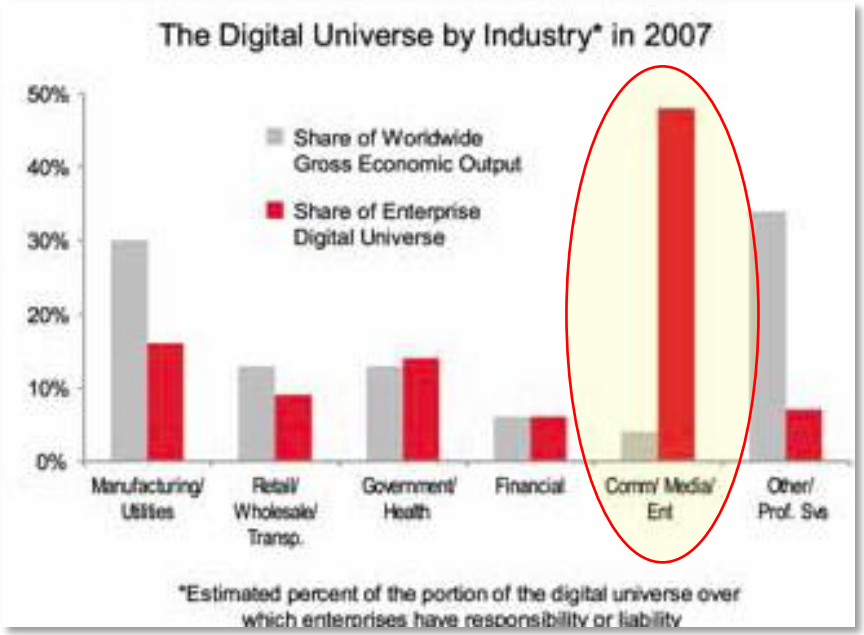
Répétitions non filtrées

Alta Plana

L’univers du contenu numérique

“La télédiffusion, les médias et les industries du loisir recueillent environ 4% des revenus globaux, mais produisent déjà, dirigent, ou supervisent 50% de l'univers numérique.”

Environ 70% de l’univers numérique est créé par les individus.



“The Diverse and Exploding Digital Universe,” (IDC, 2008)

Le défi de « l'information non-structurée »

- Les sites Web, articles de journaux et magazines, images, vidéos.
- Les blogs, forums, et médias sociaux.
- Le courrier électronique, les notes et transcriptions de centres d'appel; les conversations enregistrées.
- Les sondages, formulaires de satisfaction, demandes d'indemnité et de garantie.
- Les documents, rapports, écrits scientifiques.
- Et tout autre type de document imaginable.

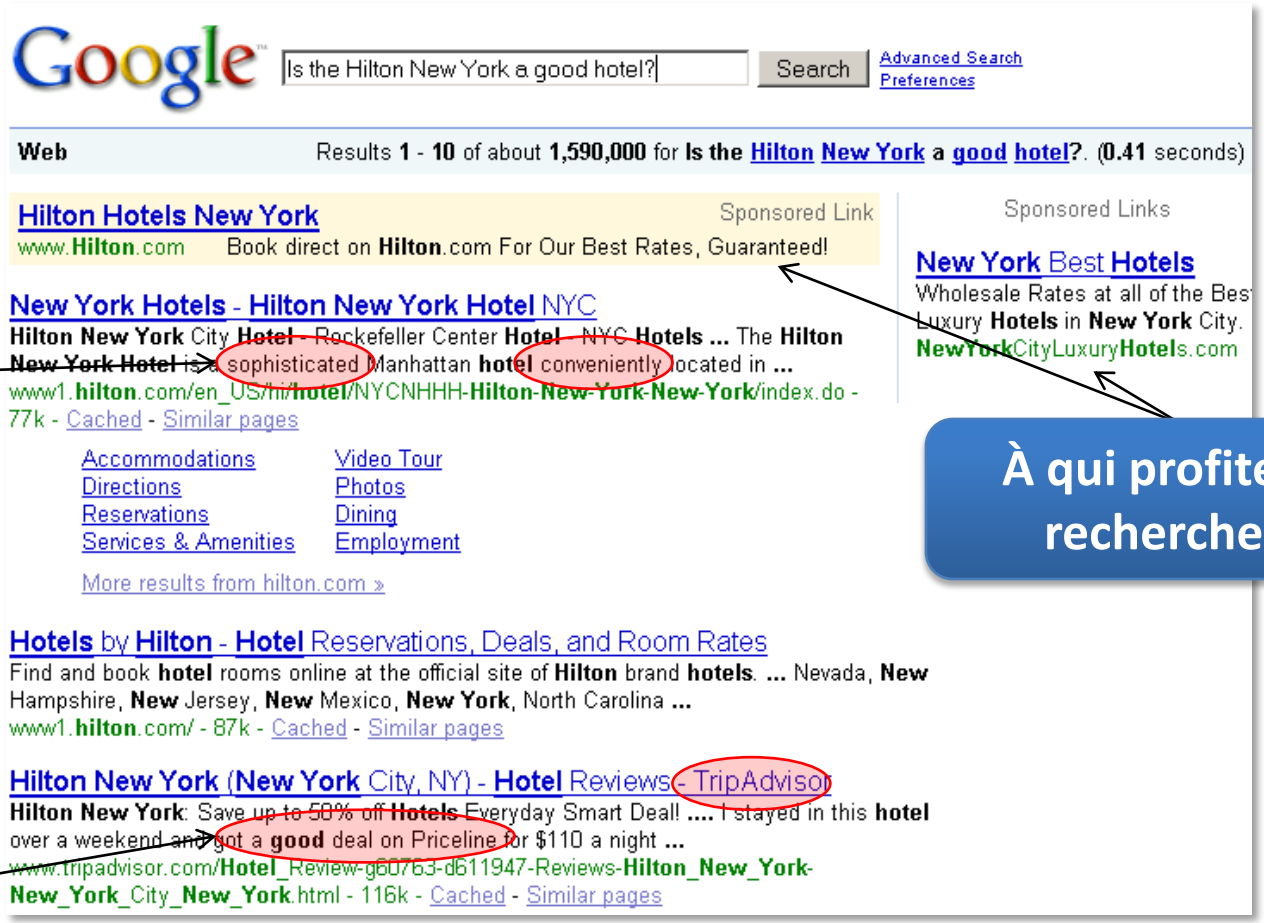
Est-ce que la recherche à elle seule est suffisante?

Résultats de recherche

Quelles sont la qualité, la valeur et la crédibilité des résultats de recherche?

L'opinion de l'hôtel

L'opinion de l'invité… à propos de Priceline



À qui profite la recherche?

Alta Plana

Du Web 1.0 au Web 2.0

Comment pouvons-nous faire mieux?

«Nous avons en place plusieurs instruments -- des technologies Web 2.0…»

"The Diverse and Exploding Digital Universe,"
(IDC, 2008)

Le Web 2.0

«Le Web 2.0 est la révolution d'affaires dans l'industrie de l'informatique provoquée par l'évolution de l'Internet comme une plate-forme.»

-- Tim O'Reilly, 2004

Le Web 2.0 marque un «mouvement des sites Internet personnels aux blogs et l'agrégation de blogs, de la publication à la participation, … un processus en cours et interactif ... vers les liens basés sur le balisage.»

-- Terry Flew, "New Media: An Introduction," 2008

Le Web 2.0 est interactif, personnalisé, collaboratif, dynamique.

Économiquement parlant

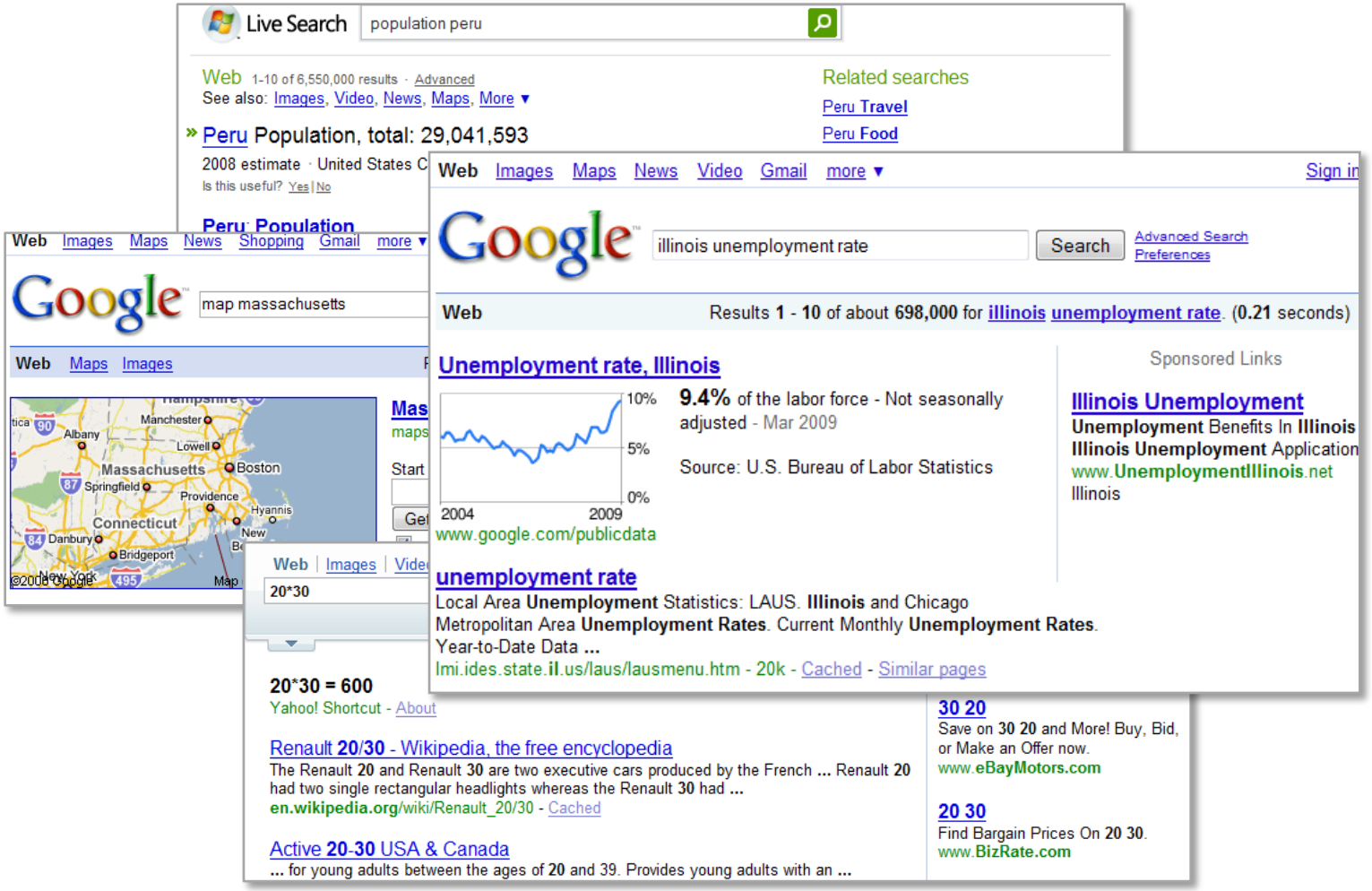
Mais comment pouvons-nous faire mieux?

«Nous avons en place plusieurs instruments -- des technologies Web 2.0… logiciels qui fouillent les données non structurées et le Web Sémantique -- pour apprivoiser l'univers numérique. **Fait de manière adéquate, nous pouvons transformer la croissance d'information en croissance économique.**»

"The Diverse and Exploding Digital Universe,"
(IDC, 2008)

Text mining: de l'information à l'intelligence

Le text mining permet une recherche plus intelligente, basée sur les objectifs de l'utilisateur, par exemple la réponse aux questions –



Du Web 2.0 au Web 3.0

Pour trouver encore plus facilement:
«Le Web sémantique est un web de données, d'une certaine manière comme une base de données globale.»

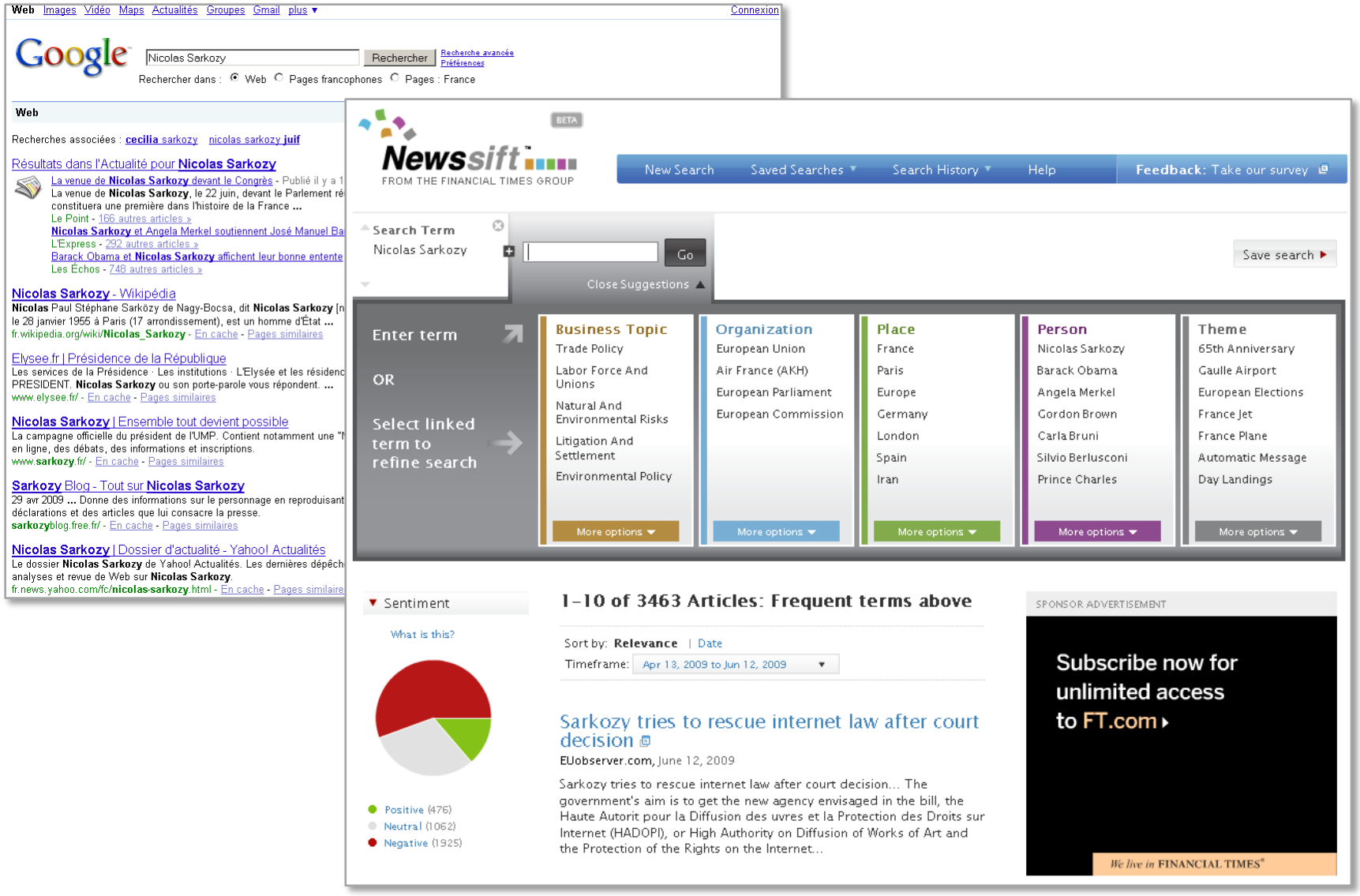
-- Tim Berners-Lee, 1998

Web 3.0 = Web 2.0 + Web sémantique + outils sémantiques.

Thèmes fréquents du Web 3.0:

- Contenus enrichis sémantiquement.
- Données reliées (« linked data »)
- Sensibilité au contexte.
- Sensibilité à la localisation géographique.

Exemples



Alta Plana

Le Web 3.0

Le text mining permet le Web 3.0 et le Web sémantique.

- La catégorisation et la classification automatique du contenu.
- L'augmentation de texte: la création de métadonnées; le balisage du contenu.
- L'extraction d'informations vers les bases de données.
- L'analyse exploratoire et la visualisation.

Concepts techniques:

- Les Microformats
- RDF, SPARQL
- OWL

Text Mining: perspective des utilisateurs

J'ai récemment publié un rapport, "Text Analytics 2009: User Perspectives on Solutions and Providers" («Text Analytics 2009: les perspectives des utilisateurs sur les solutions et les fournisseurs»).

J'ai estimé un marché global de $350 millions en 2008, une croissance de 40% par rapport à 2007.

J'ai présenté les résultats d'un sondage dans lequel j'ai posé les questions…

Principales applications

Quelles sont pour vous les applications primaires où le texte joue un rôle?

Alta Plana

Types d'informations textuelles analysées

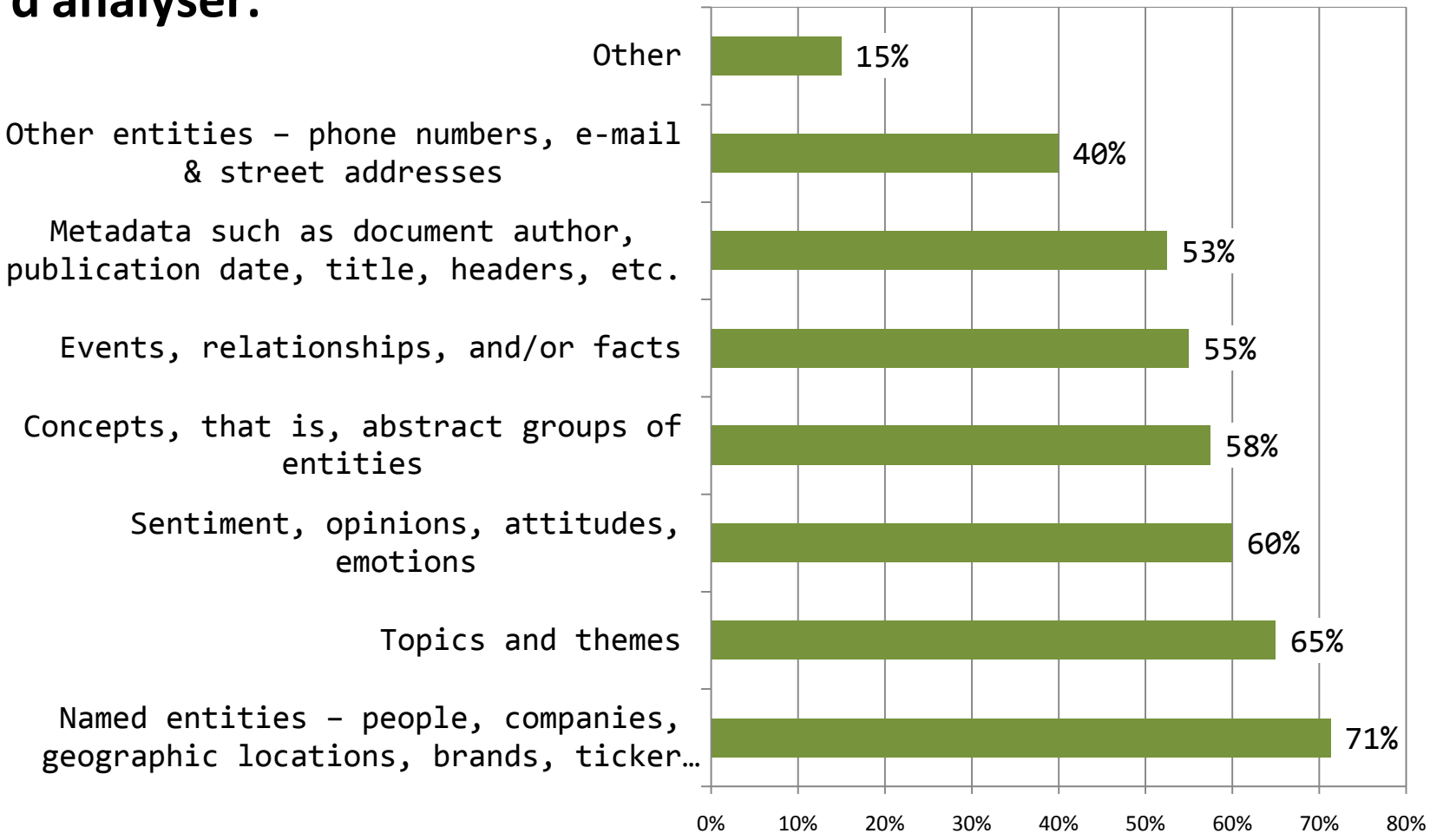
Quelles informations textuelles analysez-vous ou projetez-vous d'analyser?

Les utilisateurs **actuels** ont répondu:

blogs and other social media (twitter, social-network sites, etc.)	62%
news articles	55%
on-line forums	41%
e-mail and correspondence	38%
customer/market surveys	35%

Types d'informations extraites

Avez-vous (ou comptez-vous avoir) besoin d'extraire ou d'analyser:



Alta Plana

Satisfaction générale

Veuillez évaluer votre expérience générale – votre satisfaction – avec le text mining.