Włodzimierz Gogołek
wg@id.uw.edu.pl
Instytut Dziennikarstwa
Uniwersytet Warszawski
Warszawa

Nowy wymiar zasobów informacyjnych WWW

Wstęp

Po raz pierwszy suma cyfrowych informacji wyprodukowanych na świecie w ciągu jednego roku (2010) przekroczyła jeden zeta bajt (10^{21}). Ilość informacji produkowana każdego roku wzrasta o 40 procent [Wollan, 2011]. Stworzyło to nowe wyzwania dla nauki, edukacji, mediów i administracji. Zasoby o tej skali znane jako Big Data – ogromne nieustrukturyzowane hurtownie danych – przekroczyły krytyczną wielkość zarejestrowanych informacji. Ich ilość stworzyła nowy wymiar atrakcyjności zasobów informacyjnych do wszelkiego rodzaju badań. Krytyczne – oznacza, że konwencjonalne narzędzia do analizy tak dużych baz danych są bezużyteczne. Spowodowało to rozpoczęcie prac nad eksploracją – specjalistyczną analizą Big Data. Wyniki uzyskane z analizy dostarczyły, nigdy wcześniej niedostępne, źródła informacji. Może być to postrzegane jako nowa faza rozwoju aplikacji IT (narzędzi i cyfrowych sieci wymiany informacji).

Big Data

Ogromne nieustrukturyzowane informacje w połączeniu z łatwo skalowalnymi platformami komputerowymi urealniły cel badaczy – poszukiwanie odpowiedzi na pytania, które wcześniej pozostawały bez odpowiedzi. Umiejętna analiza Big Data może być zastosowana do doskonalenia rozwoju kolejnych generacji produktów i usług, które są wykorzystywane przez wszystkie branże – w tym przez szeroko pojętą edukację. W szczególności, doskonaląc rozumienie potrzeb edukacyjnych i zaniedbań w udostępnianym zakresie wiedzy. Owa analiza pozwoli na precyzyjniejsze, i w stosownym czasie, udostępnianie potrzebnych i krytycznych informacji.

Znaczącą część Big Data tworzą zasoby Internetu, włączając sieci społecznościowe. Dane tego typu są tworzone przez i o indywidualnych użytkownikach sieci społecznościowych (blogi, posty, portale, maile czy strumień kliknięć internetowych), profesjonalne publikacje i inne bogate zasoby informacyjne. Interesującą częścią Big Data są zasoby ukryte w Sieci (Dark Net) – Deep Web i pNet zbudowane na bazie P2P – nazywane – F2F (przyjaciel do przyjaciela), np. Freenet. Zasoby te są tysiąc razy większe od dostępnych w tradycyjnej, indeksowanej przez wyszukiwarki sieci WWW [Boswell, 2012].

Big Data to także potężne zasoby informacyjne produkowane przez telefony komórkowe (5 miliardów aparatów w 2010 roku), komputery, cyfrowe aparaty i kamery, czytniki RFID, GPS-y, samochody, także mieszkania (np. smart meters – odzwierciedlające, na podstawie intensywności używania energii elektrycznej, niektóre zachowania mieszkańców).

Zasoby zgromadzone w Big Data tworzą informacje źródłowe. Wynik ich analizy to informacje wtórne lub inaczej wynik rafinacji informacji. Przyjęto, iż proces owej analizy określany jest jako rafinacja informacji (rafinacja).

Rafinacja

Jednym z ugruntowanych już filarów rafinacji jest Culturomics. Obejmuje on aktywności związane z eksploracją kulturowych trendów poprzez analizę bogatych zbiorów umożliwiających spojrzenie na funkcjonowanie społeczeństw. Korzystanie z narzędzi Culturomicsu sprawnie sygnalizują ważne kulturalne, naukowe i historyczne zmiany [Jean-Baptiste i inni, 2011, s. 176-182]. Mając na uwadze szersze spektrum informacji – Big Data – proces uzyskiwania nowych informacji, głównie z WWW nazwano rafinacją informacji. Umożliwia ona dostrzeganie w obszarze informacji podstawowych – informacje wtórne, które są ukryte w zasobach WWW. Rafinacja jest jak mikroskop umożliwiający, jak nigdy wcześniej, zainteresowanym oglądanie i mierzenie rzeczy – na poziomie poszczególnych komórek (rekordów) jak i grup społecznych. Jest to rodzaj rewolucji w pomiarach. Uzyskane w ten sposób dane tworzą obraz potrzeb i zachowań indywidualnych użytkowników, ale także społeczności jako całości.

Narzędzia rafinacji

Używając odpowiednich narzędzi do rafinowania milionów postów, blogów i artykułów dostępnych online, jest możliwe uzyskanie wcześniej niezauważanych informacji dotyczących: społecznych fenomenów, państw, organizacji i osób indywidualnych. W wyniku rafinacji można także uzyskiwać wartościowe informacje dotyczące: oceny emocjonalnych relacji – sympatia, uraza/rozgoryczenie, poczucie szczęścia, optymizm, pesymizm, obawa, niepokój.

Wyszukiwarki są najczęściej względnie prostym (liczba funkcji) wykorzystywanym narzędziem do rafinacji informacji. Paradoksalnie wyniki wyszukiwania (SERP) są rezultatem kreacji specyficznego, zawężonego obrazu pierwotnych informacji dostępnych w Sieci. Ta specyfika bazuje na szczególnym, zamierzonym wyborze odpowiedzi na zadawane pytania. Poszukujący informacji otrzymuje odpowiedzi – SERP – które są zgodne z zadanymi kryteriami/pytaniami i partykularnym celem konstruktorów wyszukiwarki. Innymi słowy wynik wyszukiwania jest rezultatem specyficznej funkcji oczekiwań użytkownika i jednocześnie spełnienia celów (np. komercyjnych, politycznych) właściciela wyszukiwarki [Pfanner, 2010, s. 17].

Dane behawioralne stanowiące względnie nową kategorię informacji pozyskiwanych z Sieci są wynikiem potencjału znaczeniowego danych, uzyskanych z analizy sposobów przeszukiwania Sieci przez poszczególne osoby. Przykładem tego jest gromadzenie i analizowanie danych/słów/fraz używanych przez internautów podczas szukania przez nich, np. za pomocą Google'a. Dzięki temu przeprowadzane są analizy zachowań konsumentów w czasie rzeczywistym. Wynikiem tego było np. uzyskanie wartościowych i wiarygodnych informacji o ryzyku nadchodzącej epidemii grypy [Butler, 2008]. Było to efektem faktu częstszych zapytań osób, które szukały w Internecie terminów powiązanych z ich poczuciem choroby. To nowe źródło informacji może być z powodzeniem używane w innych aplikacjach, wśród nich – w aplikacjach związanych z edukacją,

identyfikując informacyjne potrzeby osób uczących się, w zakresie poziomu wiedzy, wieku, doświadczenia itp.

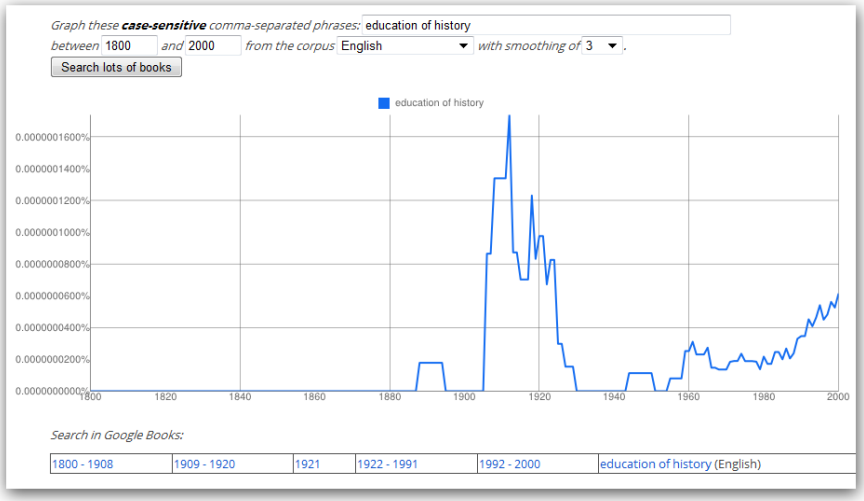
Personalizacja jest innym – bezpośrednio powiązanym z rafinacją – narzędziem. Można wyróżnić dwie przestrzenie związane z personalizacją: jedną wielowymiarową, która opisuje różne typy użytkowników (wymiary: płeć, wiek, sprawność fizyczna, zainteresowania kulturalne, sportowe itp.) oraz drugą, którą stanowią informacje pozyskiwane z Sieci odpowiednio do konkretnych wartości wymiarów opisujących użytkownika. Dzięki temu uzyskuje się, dla konkretnego użytkownika, znaczący zbiór parametrów, które zawężają (personalizują) zbiory Big Data dostępne w Sieci. W efekcie personalizacja spełnia nie tylko merytoryczne oczekiwania wirtualnego konsumenta, także jego indywidualne cechy. W rosnącej liczbie przypadków, wirtualny profil konkretnego konsumenta, np. ucznia, jest dostępny wszędzie i zawsze. Bez względu na to, gdzie i za pomocą jakiego urządzenia (PC, komórka, tablet), odbiorca ma zawsze osobiste okno dla transferu informacji dopasowanych do jego osobistego profilu. Ilustrują to bogate doświadczenia personalizowanej reklamy online, które łatwo mogą być adoptowane do potrzeb zdalnej edukacji.

Zaawansowane narzędzia. Do rafinacji zasobów sieciowych mogą być bezpośrednio użyte takie narzędzia, jak np. Attentio, Radian6, Sysomos, NetBase, Collective Intellect, Alterian, Google Alerts. Rafinacja sieciowa jest skutecznie realizowana poprzez wykorzystanie Attentio Brand Dashboard. Dowodzą tego wyniki badań dynamiki zmian obrazu informacyjnego kandydatów w wyborach prezydenckich 2010 roku [Kuczma, Gogołek, 2010, s. 35-49]. Innym profesjonalnym narzędziem rafinacji jest serwis monitorujący serwisy informacyjne – *Summary of World Broadcasts* (*SWB*). Umożliwia on monitorowanie pełnych tekstów i streszczeń artykułów prasowych, materiałów pokonferencyjnych, materiałów telewizyjnych i radiowych, periodyków i innych nieklasyfikowanych technicznych raportów w 130 językach [Leetaru, 2011].

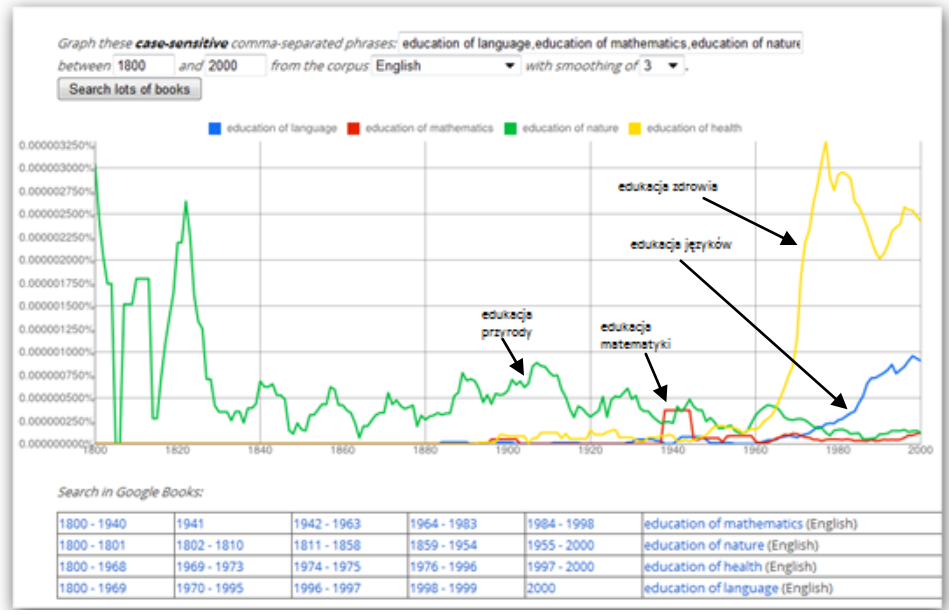
Wizualizacja jest bardzo użytecznym wymiarem rafinacji – ważnym ogniwem łańcucha łączącego źródła informacji z skutecznym wykorzystaniem rezultatów rafinacji. Problemem wizualizacji jest uzyskiwanie większej ilości informacji, ale eliminowanie mniej ważnych jej części. Wagę wizualizacji potwierdza bogactwo multimediów w mediach – włączają wizualizację statycznych i ruchomych obrazów. Reprezentatywnym przykładem takich narzędzi do rafinacji i jednocześnie ich wizualizacji są te, które są wykorzystywane do badań środowiska naturalnego – THE CARBON CAPTURE REPORT [*The Carbon*…, 2012]**.**

Wyniki rafinacji

Rafinacja stwarza możliwości wykrywania, na zadanym poziomie ufności, obrazu przeszłego i obecnego statusu informacyjnego rzeczywistości, a nawet prognozowania przyszłości. Na przykład, odnośnie do przeszłości i współczesności – korpus ponad pięciu milionów książek w formie cyfrowej umożliwia ilościową ocenę kulturalnych trendów, a używając kolektywną pamięć opublikowanych książek, rozpoznawanie adopcji technologii, cenzurę czy historię epidemiologii [Shen i inni, 2011, s. 176-182].



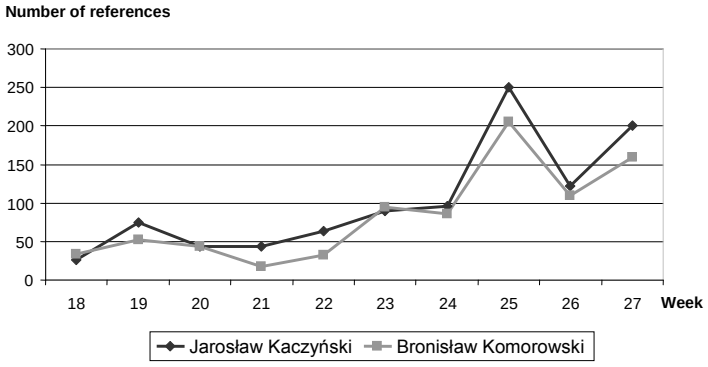
Rys. 1. Zmiany obecności zwrotu „edukacja historii" w latach 1800–2000 (liczba publikacji zawierających zwrot „Edukacja biologii, historii, sztuki" jest pomijalna)
Źródło: Wynik usługi http://books.google.com/ngrams



Rys. 2. Zmiany obecności zwrotów: edukacja matematyki, przyrody, zdrowia i języków, w latach 1800–2000
Źródło: Wynik usługi http://books.google.com/ngrams

Innym przykładem nowych danych uzyskanych dzięki rafinacji jest łączenie tradycyjnych źródeł informacji z relatywie nowymi – *crowdsourcing*. W sumie z zasobami tradycyjnymi, pozwala to uzyskać rzetelniejszy obraz o świecie, np. w zakresie doskonalenia procedur doboru przedmiotów obowiązującej edukacji (rysunki 1 i 2) [Anstey, 2012].

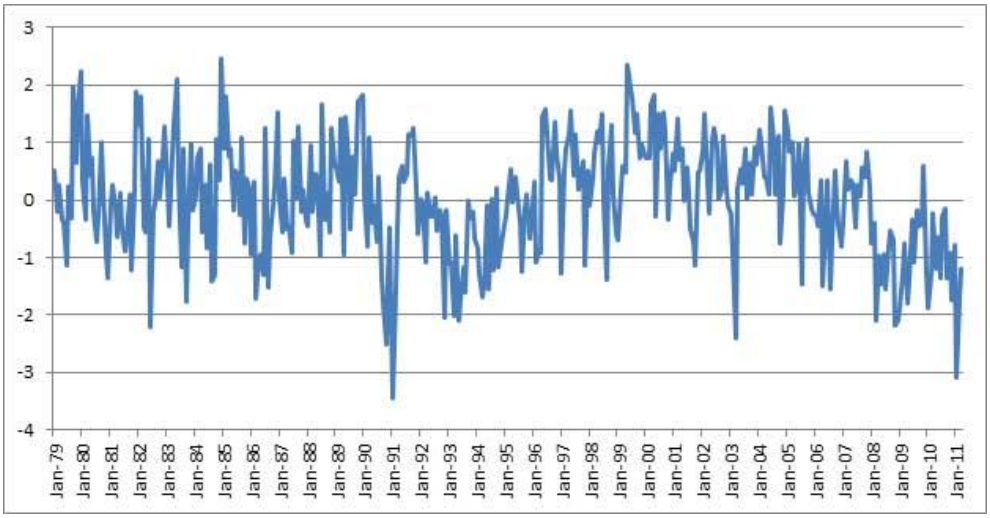
Prognozowanie wynikające z zastosowań rafinacji, ilustrują wyniki rafinacji, m.in. wspomnianego wcześniej ryzyka powstania epidemii grypy, oczekiwań zmian na rynkach finansowych, księgarskich, sprzedaży książek i wideo, a także wspomniane wcześniej badania – pozwalające przewidzieć wynik – w okresie poprzedzającym wybory prezydenckie w Polsce.



Rys. 3. Negatywne opinie zawarte w mediach społecznościowych o głównych kandydatach wyborów prezydenckich, uzyskane w dniach 5 maj–4 czerwiec [Kuczma, Gogołek, 2010, s. 35-49]

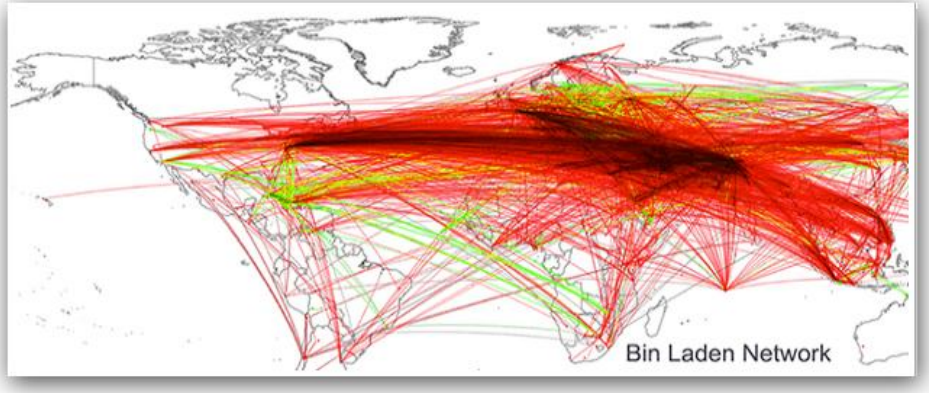
Innym znaczącym przykładem informacyjnej siły rafinacji Sieci są rezultaty badań Kalev H. Leetaru. Dokonał on analizy tonów (barw) i geograficznych wymiarów 30-letnich archiwów światowych wiadomości do konstrukcji przewidywania w rzeczywistym czasie ludzkich zachowań, takich jak narodowe konflikty i precyzyjne daty specyficznych zdarzeń [Leetaru, 2011].

Podobna problematyka dotyczyła badań tonów w skali państwa – zakresu obejmowanych w analizie informacji (rafinacja) Egiptu, Tunezji i Libii w kontekście ostatnich politycznych zmian. Rys. 4 ilustruje przeciętny ton w okresie styczeń 1979 do marca 2011, 52 438 artykułów uzyskanych z SWB, które dotyczyły Egiptu. Rysunek ukazuje zmiany tonów artykułów – pozytywne i negatywne. Odzwierciedlają one daty najbardziej istotnych zmian w tym kraju (I. 1991, III. 2003, 1-24.I.2011). Podobne rezultaty uzyskano odnośnie do dat istotnych zmian w Tunezji i Libii.



Rys. 4. Zmiany barw tonu publikacji w Egipcie [Leetaru, 2011].

Ocena barwy tonu informacji dotyczących części świata wyróżnia zapalne regiony kontynentów i innych informacji dotyczących indywidualnych postaci, np. odnośnie lokalizacji pobytu Bin Ladena. Rys. 5 pokazuje wszystkie geograficzne wskazówki dotyczące Bin Ladena zawarte w treściach SWB od stycznia 1979 do kwietnia 2011. Niemal połowa materiałów dotyczących Bin Ladena mówiła o Pakistanie [Leetaru, 2011].



Rys. 5. Globalny geolokalizacyjny ton treści Summary of World Broadcasts, styczeń 1979–kwiecień 2011 zawierających zwrot: „Bin Laden” [Leetaru, 2011].

Zakończenie

Rafinacja zasobów Big Data umożliwia ilościową analizę szerokiego spectrum pierwotnej, nieustrukturyzowanej oryginalnej informacji w celu odkrycia/wskazania znaczących problemów ludzkości – społecznych, kulturalnych, politycznych, biznesowych także związanych z szeroko rozumianą edukacją. Rafinacja kreuje nową przestrzeń wartościowych źródeł informacji i otwiera nowe drogi do badań. Narzędzie to zapewne spowoduje przełomowe zmiany w świecie informacji.

Zarządzanie informacjami i ich analiza są krytyczne w osiąganiu ważnych celów poprawnej i skutecznej edukacji. Problem nie leży w wielkości Big Data, ale w tym, że większość użytkowników nie dysponuje odpowiednią platformą narzędzi do ukierunkowanej analizy. Big Data pozostają *terra incognita* dla edukacji. Ważnym wyzwaniem jest opracowanie stosownych narzędzi do rafinacji informacji, spolegliwego dostarczania wyników użytkownikom, a co najważniejsze przekonania o użyteczności tego nowego źródła informacji [Beck, 2012].

Bez możliwości integracji Big Data z poprawną rafinacją dla edukacji, edukacja traci szansę na przyspieszone uzyskiwanie wartościowych informacji o rzeczywistym obrazie oraz potrzebach współczesnej edukacji. Informacji ułatwiających unikanie nie merytorycznych, koniunkturalnych zmian w systemach i treściach nauczania. Nauczyciele, głównie ich decydenci, którzy jako pierwsi skorzystają z potencjału rafinacyjnego Big Data będą najbliżej bieżących potrzeb uczących się, rynku i współczesnych oczekiwań społecznych.

Literatura

Anstey C.: *Empowering Citizen Cartographers.* "The New York Times"*,* Jan. 13, 2012

Beck A.: *Big Data Is Never Too Big When You Can Act On It*. http://www.clickz.com/print_article/clickz/column/2171482/act?wt.mc_ev=click&WT.tsrc=Email&utm_term=&utm_content=Print%20version&utm_campaign=05%2F02%2F12%20-%20Behavioral%20Marketing&utm_source=ClickZ%20Media&utm_medium=Email [maj 2012]

Boswell W.: *Five Search Engines You Can Use to Search the Deep Web.* http://websearch.about.com/od/invisibleweb/tp/deep-web-search-engines.htm [marzec 2012]

Butler D.: *Web data predict flu.* "Nature" 2008, Vol. 456 (287-288) http://www.nature.com/news/2008/081119/full/456287a.html [luty 2012]

Leetaru K.: *Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space*. "First Monday"September 2011, Vol. 16, Number 9-5 http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [styczeń 2012]

Jean-Baptiste M., Shen Y., Aiden A., Veres A., Gray M., Pickett J., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M., Aiden E.: *Quantitative analysis of culture using millions of digitized books.* "Science" 2011, Vol. 331, number 6014. http://www.sciencemag.org/content/331/6014/176 [czerwiec 2011]; http://www.culturomics.org/cultural-observatory-at-harvard [kwiecień 2012]

Kuczma P., Gogołek W.: *Informacyjny potencjał sieci – na przykładzie wyborów prezydenckich 2010.* „Studia Medioznawcze" 2010, nr 4(43), Instytut Dziennikarstwa UW, Warszawa 2010

Pfanner E.: *Google, in Settlement, Changes Ad Rules in France.* "The New York Times" 2010*,* October 28

Deacon D.: *Yesterday's Paper and Today's Technology, Digital Newspapers Archives and "Push Button" Content Analysis*, „European Journal of Communication" 2007, vol. 22, nr 1

Shen M., Aiden A. P., Veres A., Gray M. K., Pickett J. P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M. A., Aiden E. L.: *Quantitative Analysis of Culture Using Millions of Digitized Books*. "Science" 2011, Vol. 331, Issue 6014

The Carbon Capture Report. http://www.carboncapturereport.org/ [wiosna 2012]

Wollan M.: *For Start-Ups That Aim at Giants, Sorting the Data Cloud Is the Next Big Thing.* "The New York Times" 2011, December 25