Börteçin Ege

Hacettepe Üniversitesi
Semantik Web Doktora Öğrencisi

Yeni Nesil Arama Motorları

İnternetteki bilgi hacmi gün geçtikçe, hem de katlanarak artıyor. Günümüzün klasik arama motorlarında yapılan arama sonuçları bunun en belirgin örneği. Yaptığınız herhangi bir aramaya milyonlarca cevap almak işten bile değil, tabi bunlar cevap olarak adlandırılabilirse. Sonuç olarak Web 'deki veri yığınları içinde bilgi aramak artık samanlıkta iğne aramaya benziyor. Web'deki bu bilgi kirliliği üzerine yıllardan beri kafa yoran Google, Yahoo ve Microsoft gibi bilişim devleri bu konuya bir çare bulabilmek için çalışıyor. Söz konusu çalışmalarda henüz istenilen noktaya gelinmiş olmasa da sonuçlar hayli ümit verici. Anlamsal teknolojilerin uygulanmasıyla hayata geçirilmeye başlanan bu süreç aynı zamanda Web'de yakın bir zamanda meydana gelecek büyük bir teknolojik devrimin de habercisi.

Klasik arama motorları ve Google

Birçok bilgisayar kullanıcısı için Web arama motorlarının tarihçesi Google ile başlıyor, oysa bu doğru değil. Google her ne kadar günümüzdeki arama motorları içinde en göze çarpanı ve en başarılısı olsa da Lycos (1994), AltaVista (1995), Yahoo (1995) ve Yandex (1997) gibi arama motorlarının faaliyete geçiş tarihleri Google'ınkinden (1998) daha önce. Yani arama motorlarının tarihçesinin Google ile başlamadığını bir daha hatırlatmakta fayda var. Peki, nedir Google'ı Google yapan? Nasıl oldu da Google "Web'de arama yapmak" ve "arama motoru" kavramlarıyla özdeşleşti? Bunun cevabı esasında hayli basit ve öncelikli olarak sistemin çekirdeğini oluşturan PageRank™ algoritmasında yatıyor. Günümüzde belki Coca Cola'nın formülünden bile daha değerli olan PageRank™ algoritması, 1996-1997 yılları arasında Stanford Üniversitesi'ndeki bir doktora çalışması kapsamında Lawrence Page ve Sergey Brin tarafından geliştirildi. Algoritmanın patentinin 1997 yılında Stanford Üniversitesi tarafından alınmasından sonra, 4 Eylül 1998'de algoritmanın mucitleri Lawrence Page ve Sergey Brin tarafından özel bir şirket olarak kurulan Google, 19 Ağustos 2004'te Google Inc. adını alarak halka arz edildi. Çalışma şekli ve diğer arama motorlarından daha başarılı arama sonuçları dikkate alındığında Google haklı olarak günümüzün en iyi arama motoru unvanına sahip. Bu nedenle yazımızda Web 2.0 çerçevesinde klasik arama motorlarını değerlendirirken ölçüt olarak Google'ı alacağız.

PageRank™ algoritması

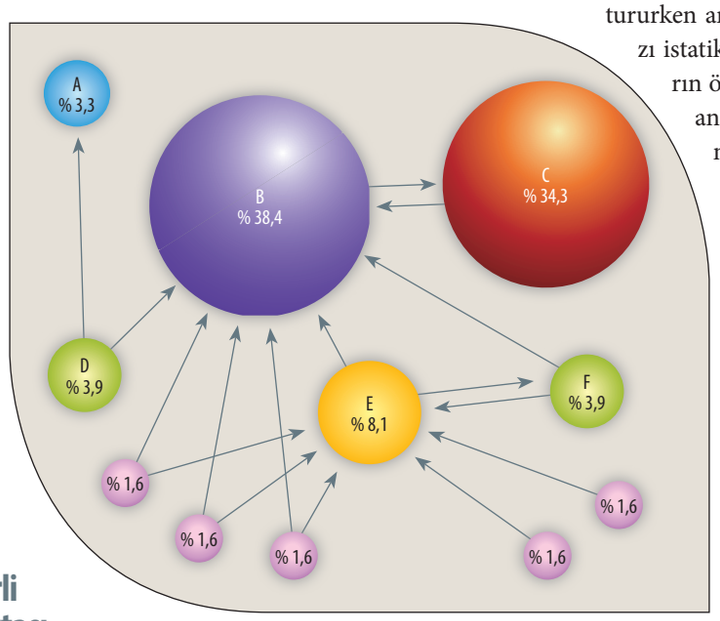
Her arama motorunun olduğu gibi Google'ın da arama sonuçlarını etkileyen çeşitli ölçütler var. Bu ölçütlerden bilinenler arasında en önemlileri şunlar: Aranan anahtar kelimeler ile bunların olası kombinasyonlarının o anda incelenen Web sayfası içindeki sayısı ile bunların bulunduğu konumlar, genel anlamda içerik uyuşması, söz konusu sayfa içinde verilen kaynakçaların güvenilirliği ve Web sitesinin Google tarafından tayin edilmiş PageRank™ değeri. Her ne kadar Google, özellikle kullanıcı tarafından aranan ifadeler ile söz konusu Web sayfasının içeriğinin uyuşması durumunun, sonuçların sıralanmasına çok büyük etkisi olduğunu belirtse de, bir değerlendirme sürecine etki eden fakat Google tarafından özellikle açıklanmayan başka faktörler de var. (Google'ın, PageRank™ değerini hesaplayan algoritmasını yeni gereksinimlere göre sürekli güncellediği, fakat bu değere etki eden tüm faktörleri kamuoyuyla paylaşmadığı bilinen bir gerçek. Bunun en önemli nedenlerinden biri PageRank™ algoritmasının müthiş bir ticari sır olmasının yanı sıra, Google'ın birtakım işgüzar ağ yöneticileri tarafından gerçekleştirilebilecek yapay arama motoru optimizasyonlarının önüne geçerek PageRank™ algoritmasının bu tipteki sahte bağlantılarla ve değerlerle yanıltılmasını önlemek istemesi.

>>>

PageRank™ algoritmasının tarihçesi

PageRank™ algoritmasının kökleri sosyometri bilimine dayanıyor. Temelleri 1930'lu yıllarda, Avusturya kökenli ABD'li bilim adamı Jacob Levy Moreno tarafından atılan sosyometri biliminin ana fikri bir grubun üyeleri arasındaki ilişkilerin, sosyomatriks olarak adlandırılan bir matriksin yardımıyla tespit edilmesi, ortaya çıkan sonuç tablosunun sosyogram adı verilen bir grafikle görselleştirilmesidir. Sosyometri bilimi günümüzde sosyal ağların analiz edilmesinde kullanılan bilimsel yöntemlerin de babası olarak kabul edilir.

Sosyometri biliminden esinlenilerek geliştirilen PageRank™ algoritmasının temel ilkesine göre, ne kadar çok Web sayfası belirli bir Web sayfasına referansta bulunup o sayfayı işaret ediyorsa, o Web sayfasının "ağırlığı" dolayısıyla PageRank™ değeri o kadar yüksek olur (PageRank™ algoritmasının çalışma ilkesine göre herhangi bir Web sayfasının PageRank™ değeri 0 ile 10 arasında olabilir).



Bir ağda yüzdelerle ifade edilmiş PageRank değerleri.
C sayfasını gösteren bağlantı sayısının daha az olmasına rağmen, C sayfasının değeri E sayfasınınkinden daha fazla.
Bunun nedeni C sayfasını gösteren bağlantının, "çok önemli" olarak kategorize edilen B sayfasından gelmesi.

Günümüzün en değerli formülünün zayıf noktası

Yüz milyonlarca Web sayfasının arama sonuçlarında sürekli en başlarda çıkmak için kıyasıya mücadele verdiği günümüzde, PageRank™ algoritmasının değerinin Coca Cola'nın formülününkini çoktan geride bıraktığını söylersek herhalde abartmış olmayız. Fakat kelimenin gerçek anlamıyla Google'ın kalbini teşkil eden bu PageRank™ algoritmasının Web sayfalarının puanlarını tayin ederken Web kullanıcılarına adaletli davrandığı da söylenemez. Nitekim, kullanıcı tarafından yapılan bir aramanın sonuçları PageRank™ algoritması tarafından derlenirken -içerik uyuşması açısından zayıf da olsa- ilk sıralarda sadece PageRank™ değeri yüksek olan Web sayfalarına yer verilmesi, aslında söz konusu aramaya daha iyi hatta en iyi cevabı verdiği halde bazı Web sayfalarının -sadece PageRank™ değeri daha düşük olduğu için- çok daha alt sıralarda yer alabilmesi, kullanıcıların günlük hayatta çok sık karşılaştığı rahatsız edici gerçeklerden biri. Bilgilerin sadece insanlar tarafından anlaşılabilir metinsel bir formatta saklanabildiği, metinsel bazlı Web sitesi sayısının dünya genelinde yaklaşık 650 milyona ulaştığı Web 2.0 ortamında, artık PageRank™ algoritmasının tek başına gittikçe çaresiz ve yetersiz kalmaya, hatta bu şekliyle geçerliliğini yitirmeye başladığını kolaylıkla söyleyebiliriz.

Web 2.0'ın yapısal problemleri

Bilindiği gibi Web'in ilk nesli olan Web 1.0 (1995-2000) yalnızca HTML belgelerin yer alabildiği "donuk" bir yapıya sahipti. Web 2.0 (2000-2010) ile birlikte kullanıcıların da aktif olarak katılabildiği etkileşimli ve insan odaklı bir platform doğdu. Facebook, Twitter, YouTube gibi günümüzün en popüler ve önemli kitlesel iletişim araçları da Web 2.0 sayesinde doğdu ve dünyamıza kelimenin tam anlamıyla yeni bir dinamizm geldi.

Sağladığı tüm olanaklara rağmen, sonuç olarak Web 2.0 da metin tabanlı. Bu da, Web'deki bilgi hacminin her geçen gün katlanarak arttığı günümüzde büyük problemlere yol açıyor. Bunların en başında da metinsel bilgilerin sadece insanlar tarafından anlaşılması, bilgisayarlar tarafından anlaşılamaması geliyor (Web'deki içerikler metinsel tabanlı olarak saklandığı sürece de, bu mümkün olacak gibi görünmüyor). Böyle bir ortamda Google gibi en başarılı arama motorları bile, arama sonuçlarını oluştururken anlamsal analizlerden çok, bazı istatiksel verilerin ve hesaplamaların ön planda tutulduğu niceliksel analizlerle hareket ediyor ve sonuçta kullanıcı için pek de verimli olmuyorlar.

Web 2.0'ın yapısal problemleri şu şekilde özetlenebilir:

- Bilgilerin büyük bir kısmının metinsel kaynaklı ve dolayısıyla yalnızca insanlar tarafından anlaşılan yapıda olması
- Bilgilerin büyük bir kısmının metinsel kaynaklı olmasının, bu bilgilerin anlamlandırılıp bilgisayarlar tarafından "anlaşılmasını" ve aralarında gerekli bağlantıların kurulmasını engellemesi
- Bilgiler "anlamlandırılamadığı" için eş anlamlılık, çok anlamlılık gibi problemlerin çözümlenememesi ve bunun doğal olarak arama sonuçlarına yansıması
- Aralarında bağlantı kurulamayan bilgilerden yeni bilgi çıkarsamanın imkânsız hale gelmesi
- Web'de bulunan belge içeriklerinin bilgisayar tarafından anlaşılamamasının, günümüz Web'inin hemen hemen hiç bir kontrolün ve dolayısıyla sanal güvenliğin olmadığı bir ortam haline gelmesine ister istemez katkıda bulunması

Yani çok geç olmadan insanlığın Web'in içeriğine hâkim olması gerekiyor. Bunun yolu ise Web 2.0'da yaşanan problemlerin yaşanmayacağı, yeni nesil teknolojilerle modellenecek yeni nesil bir Web'den geçiyor. Bunun adı ister Web 3.0, ister Web of Data, isterse Semantik Web olsun.

Web'in geleceği: Semantik Web

Esasında olayların bu şekilde gelişeceği daha doğrusu gelişmesi gerektiği daha 2000'li yılların başında bir grup bilim insanı tarafından öngörülmüştü. James Hendler, Ora Lassila ve Web'in mucidi Tim Berners-Lee önderliğinde 17 Mayıs 2001'de *The Scientific American*'da yayımlanan "The Semantic Web" başlıklı yazıda, yazarlar ilk defa içeriği sadece insanlar tarafından değil, bilgisayarlar tarafından da anlaşılacak yeni nesil bir Web yani Semantik Web fikrini ortaya atıyordu. Yine aynı makalede Semantik Web'in çeşitli anlamsal teknolojilerin kullanımıyla nasıl hayata geçirilebileceğini açıklayan bu bilim insanları, aynı zamanda geriye dönüşü olmayan bir dönemin başlangıcını da duyuruyordu.

Bilişim dünyası o tarihten bu yana Semantik Web konusunda çok mesafe aldığı için şanslıyız. Kullanıcılar henüz Web'in daha çok alt katmanlarında gerçekleşen bu sessiz devrimden haberdar olmayabilir, ama daha şimdiden arama sonuçlarındaki bazı iyileşmeleri anlamsal teknolojilerin günümüz Web'ine yavaş ama emin bir şekilde entegre edilmesine borçluyuz. Bu şekilde uzun vadede Web'de bulunan her bilgiye ilgili anlamın yüklenmesi, daha sonra bu bilgilerin birbirleriyle ilişkilendirilip birbirine "bağlanması", böylece bütün Web 2.0'ın içeriklerden çok içeriklerin anlamının ön planda tutulduğu Semantik Web'e, sonuç olarak da küresel ölçekte "akıllı" bir veri tabanına dönüştürülmesi tasarlanıyor.

"Türkiye'nin başkenti neresidir?" sorusuna Hakia'nın Google'ınkinden hiç de aşağı kalmayan cevapları

Bilişim devlerinin "anlamlı" hayalleri

Yukarıda da belirtildiği gibi Web 2.0 ortamına hâkim olan belirsizlik aslında Google, Yahoo ve Microsoft gibi diğer bilişim devlerini de uzun zamandır huzursuz ediyor olmalı ki, özellikle son yıllarda Web'in dolayısıyla arama motorlarının bu ana problemini çözmek için çalışmalarını hızlandırdılar. Şimdi gelin bu çalışmalara bir göz atalım.

Hakia: 2004'te ABD'de Türk bilim adamı Dr. Rıza C. Berkan önderliğinde geliştirilmeye başlanan Hakia aynı zamanda ilk anlamsal arama motorlarından biri. Kuruluşundan bu yana özellikle yapay zekâ, bulanık mantık, dil bilimleri ve özellikle anlamsal teknolojilerin kullanımıyla QDEX™ (*Query Detection and Extraction*) ve SemanticRank™ gibi hayli yenilikçi algoritmaların geliştirilmesine imza atan Berkan, halen ABD'nin New York kentinde çalışmalarına başarıyla devam ediyor.

Hakia, 2008'de en iyi 10 Semantic Web ürününden biri seçilmişti.

Powerset (Microsoft Bing): 2005'te ABD'nin San Francisco kentinde eski NASA ve Xerox PARC çalışanlarının da aralarında bulunduğu bir grup tarafından kurulan Powerset'in vizyonu ve amacı, Web'de yazılan her cümleyi anlamaktı. PARC (*Palo Alto Research Center*) araştırma merkezinden bir doğal dil işleme teknolojisi satın alarak yola çıkan Powerset, başarılı çalışmalar sonucunda gelecek vaat eden bir anlamsal arama motoru haline geldi. 2008'de de Microsoft tarafından 100 milyon ABD doları karşılığında satın alındı.

2008'den itibaren yayın hayatına Microsoft Bing adıyla devam eden Powerset, yine 2008'de en iyi 10 Semantik Web ürününden biri seçilmişti.

DBpedia: 2007'de yayın hayatına başlayan DBpedia, her ne kadar kullanıcının anladığı türden tipik bir anlamsal arama motorunu temsil etmese de, geleceğin anlamsal arama motorlarına, tüm bilgilerin Semantik Web formatında sunulduğu çok geniş kapsamlı bir bilgi kümesi sağlaması açısından hayli önemli. İki Alman üniversitesi (Freie Universität Berlin ve Universität Leipzig) ile OpenLink Software firması tarafından tasarlanan DBpedia kelimenin gerçek anlamıyla türünün en nadir ve başarılı örneklerinden biri.

DBpedia projesinin özünde, İnternet ansiklopedisi Wikipedia'daki farklı dillerdeki bilgilerin -anlamsal teknolojilerin kullanımıyla geliştirilmiş algoritmaların yardımıyla- Semantik Web formatındaki bilgilere dönüştürülmesi yatıyor (bir sonraki aşamada ise mümkünse bu bilgiler GeoNames, Freebase gibi yine Semantik Web formatındaki diğer bilgi kümeleriyle birbirine "bağlanarak" Linked Open Data Bulutu -*LOD Cloud*- kapsamında yayımlanıyor). DBpedia, özellikle de geliştirilme aşamasındaki anlamsal arama motorları için hayli faydalı ve geniş kapsamlı bir bilgi kümesi sunmasının yanı sıra, RelFinder (*Interactive Relationship Discovery in RDF Data*) adlı bir araç sayesinde normal Web kullanıcıları tarafından da etkileşimli olarak sorgulanabiliyor.

Birçok Semantik Web projesine ilham kaynağı olan ve yayın hayatına başarıyla devam eden DBpedia, 2009'da en iyi 10 Semantik Web ürününden biri seçilmişti.

Yahoo SearchMonkey: 2008'de Yahoo tarafından Semantik Web formatındaki yapısal bilgilerin kullanımının araştırılması ve anlamsal arama amaçlı testlerin yapılması için kurulan SearchMonkey adlı arama motoru da, anlamsal arama yapabilen ilk arama motorlarındandı. Yaklaşık iki buçuk yıl hizmette kalan SearchMonkey, Ekim 2010'da yine Yahoo tarafından hizmetten kaldırıldı.

SearchMonkey, 2008'de en iyi 10 Semantik Web ürününden biri seçilmişti.

WolframAlpha: Dünyaca ünlü matematiksel yazılım Mathematica'nın mucidi Stephen Wolfram'ın önderliğinde 2005'te geliştirilmeye başlanan WolframAlpha, yukarıda tanıtılan anlamsal arama motorlarından biraz daha farklı olarak aranılan bilgiyi sadece Web'de aramakla kalmaz, bulduğu veriler arasında bağlantı kurarak, mümkünse yeni sonuçlar da çıkarır.

WolframAlpha'ya arama ifadesi olarak örneğin bir ülke veya kent adı girildiğinde, hayli detaylı bilgi ve özellikle matematiksel türdeki işlemlerde de hayli kesin sonuçlar alınır. Örneğin kullanıcının WolframAlpha'da "İstanbul" kelimesini araması sonucunda, kullanıcıya İstanbul'un haritadaki yeri, koordinatları, kuruluş tarihi, o andaki hava durumu, yerel saat, kent merkezine en yakın havaalanı gibi önemli ayrıntılar gösterilir.

2009'da yayın hayatına başlayan WolframAlpha günümüzde hâlâ kendi kategorisinin en başarılı anlamsal arama motorlarından biri olarak görülüyor.

Google: Mart 2012'de Google'ın başmühendislerinden Amit Singhal, ABD'nin ünlü *The Wall Street Journal* gazetesine verdiği röportajda, çok yakında bugünkü klasik arama sistemini -yine Google tarafından geliştirilen bazı anlamsal arama algoritmalarıyla bütünleştirerek- iyileştirmeyi planladıklarını açıkladı. Geliştirmekte oldukları bu doğal dil işleme tabanlı algoritmalar sayesinde, Google'ın karşılaştığı her kelimenin gerçek anlamını zorlanmadan çözeceğini belirten Singhal, yeni sistemin milyonlarca Web sitesinin sıralamasına da büyük etkisi olacağını belirtti. Bu sürecin aynı zamanda Google'ın tarihinde yaşayacağı en büyük değişikliklerden biri olacağını ifade eden Singhal, bunun nasıl gerçekleştirileceği konusunda detaylı bilgi vermedi.

Anlamsal arama motorlarında bazı sorgulama örnekleri: Anlamsal arama motorlarının kullanıcılar açısından en büyük faydalarından biri, kullanıcıların sorularını anahtar kelimeler yerine soru cümlesi şeklinde sormasını sağlamak olacak. Kullanıcı tarafından girilen sorguyu "anlayacak" derecede gelişmiş olacak bu sistem, sadece sorudaki tek tek kelimelerin ne anlama geldiğini gerçekten kavramakla kalmayacak aynı zamanda sorulan sorunun bütün olarak ne anlama geldiğini de "bilecek". Örneğin ABD Başkanı Barack Hussein Obama'nın ne zaman doğduğunu öğrenmek isteyen bir kullanıcı, arama motoruna "Barack Obama" + "doğmak" gibi anahtar kelimeler girmek yerine, doğrudan "Barack Obama ne zaman doğdu?" diye yazabilecek ve sorusunun karşılığı olarak da "4 Ağustos 1961" veya "Barack Hussein Obama 4 Ağustos 1961'de doğmuştur" gibi bir cevap alacak.

Anlamsal arama motorlarının başka bir kuvvetli yönü de eş anlamlı ve çok anlamlı ifadeleri aynı insanlar gibi ayırt edebilecek olması. Örneğin kullanıcının arama motoruna "Doğan model bir araba satın almak istiyorum" şeklinde bir soru yöneltmesi durumunda, arama motoru "Doğan" sözcüğü ile bir yırtıcı kuşun ya da "doğmak" ile ilgili bir kavramın kast edilmediğini otomatik olarak anlayacak ve arama sürecini satışa çıkarılmış Doğan model arabaların bulunmasına göre yönlendirecek.

Sonuç

Görüldüğü gibi bilişim devlerinin Web'deki bilgi kirliliği ile mücadelesi her geçen gün hem kullanıcılar hem de bilişim devleri açısından gittikçe önem kazanıyor. Bu mücadelenin kazanılması ise ancak Semantik Web kavramının hak ettiği yere tam anlamıyla gelmesiyle mümkün. Fakat anlamsal arama motorlarının önünde duran en önemli engel teknoloji değil, bilgilerin metinsel formatta saklandığı Web 2.0'ın ta kendisi. Her ne kadar anlamsal arama motorlarının geliştirilmesi konusundaki ilerlemeler ve bilişim devlerinin bu konudaki gayretleri ümit verici olsa da, Doğal Dil İşleme (*Natural Language Processing*, kısaca *NLP*) yani içeriğin bilgisayarlar tarafından anlaşılması konusundaki bazı problemler -ağırlıklı olarak da Web 2.0'ın metinsel yapısından dolayı- devam ediyor (bu noktada Web 2.0'daki bilgilerin programlanmış bazı algoritmalar yoluyla yani insan eliyle anlamsal formata dönüştürülmesi ise -Web'in hacmi düşünüldüğünde- pek de üstesinden gelinebilecek bir iş gibi görünmüyor).

Anlamsal arama motorlarının artık Linked Open Data Bulutu gibi sistemler dışında da başarıyla kullanılabilmesi için Web'de yayımlanan bilgilerden anlamsal formatta olanların oranının artırılması, metinsel bazda olanlarının oranının da azaltılması gerekiyor. Bu konudaki en önemli görev ise tahmin edilebileceği gibi kullanıcılara düşüyor.

Sonuç olarak içeriğini sadece insanların değil aynı zamanda bilgisayarların da "anladığı" bir Web doğduğu gün, o Web artık küresel çapta geçerli bir anlamsal veri tabanına dönüşecek ve insanoğlu bir zamanlar açık denizlere hâkimiyet konusunda kazandığı zafer gibi en önemli zaferlerinden birini kazanmış sayılacak.

Kaynaklar
Berners-Lee, T., Hendler, J. ve Lassila, O., "The Semantic Web", *Scientific American*, 17 Mayıs 2001.
Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. ve Hellmann, S., "DBpedia - A Crystallization Point for the Web of Data", *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Cilt 7, Sayı 3, s. 154-165, Eylül 2009.
Helft, M., "In a Search Refinement, a Chance to Rival Google", *The New York Times*, 9 Şubat 2007.
NETCRAFT, "Web Server Survey", **http://news.netcraft.com/archives/2012/03/05/march-2012-web-server-survey.html**, Mart 2012
Internet World Stats, "World Internet Usage and Population Statistics", **http://www.internetworldstats.com/stats.htm**, 30 Haziran 2012
ReadWrite, "Top 10 Semantic Web Product of 2008", **http://readwrite.com/2008/12/02/top_10_semantic_web_products_2008**, 2 Aralık 2008
ReadWrite, "Top 10 Semantic Web Product of 2009", **http://readwrite.com/2008/12/02/top_10_semantic_web_products_2009**, 2 Aralık 2009