EGE ÜNİVERSİTESİ

(DOKTORA TEZİ)

SEMANTİK BİLGİNİN ANALİZİ VE MODELLENMESİ

Selma TEKİR

Tez Danışmanı : Prof. Dr. Şaban EREN

İkinci Danışmanı : Doç. Dr. Ahmet KOLTUKSUZ

Bilgisayar Mühendisliği Anabilim Dalı

Bilim Dalı Kodu : 619.01.00
Sunuş Tarihi : 08.10.2010

Bornova-İZMİR

2010

Sayın **Selma TEKİR** tarafından DOKTORA TEZİ olarak sunulan "**Semantik Bilginin Analizi ve Modellenmesi**" başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 08.10.2010 tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:		**İmza**
Jüri Başkanı	**: Prof. Dr. Şaban EREN**	**.....**
Raportör Üye	**: Doç. Dr. Ahmet KOLTUKSUZ**	**.....**
Üye	**: Yrd. Doç. Dr. Timur KÖSE**	**.....**
Üye	**: Prof. Dr. İsmet KARACA**	**.....**
Üye	**: Yrd. Doç. Dr. Serap ATAY**	**.....**

ÖZET

SEMANTİK BİLGİNİN ANALİZİ VE MODELLENMESİ

TEKİR, Selma

Doktora Tezi, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Şaban EREN
Ekim 2010, 94 sayfa

Bilgi, insanın nitel ve nicel gözlemler sonucunda doğrudan elde ettiği veri ve/veya bu veriyi işleyerek, analiz ederek dolaylı ve/veya dolaysız olarak ulaştığı sonuçlar ve kanılardır. Yapılan analizin niteliği, bilginin soyutlama seviyesini belirler. Semantik analiz, doğrudan ölçülebilen bilgi niteliklerinden dolaylı, daha üst seviyede yer alan bilgi niteliklerine erişilmesini sağlayan yöntemlerdir ve semantik analiz sonucunda erişilen bilgi de semantik bilgidir. Semantik analiz bir yöntemler bütünüdür, bu alanda tek, genel geçer bir yaklaşımdan bahsedilemez.

Bu tez çalışmasının amacı, bilgiyi semantik seviyede temsil edecek bir bilgi modelinin altyapısı hakkında bir değerlendirme yapmaktır.

Semantik bilgi bir manifold üzerinde modellenebilir. Manifold, varolan modellerde kullanılan metrik uzaylardan daha genel bir çerçeve sunarken, konum özelleştirmesi yolu ile özelleşmiş bilgilerin de aynı zamanda karşılanmasını sağlar. Spesifik olarak, uzunluk metriğine ek olarak eğrilik metriğinin kullanımını mümkün hale getirir.

Çalışma kapsamında, bilginin manifold üzerinde modellenmesi fikrini doğrulamak üzere Öklidyen olmayan geometri temelli bir semantik analiz yöntemi önerilmektedir. Önerilen yaklaşım manifold üzerindeki jeodezik uzunluk kavramına dayanmaktadır. Yapılan uygulamada jeodezik uzunluklar dokümanlar arasındaki benzerliği ölçmek üzere kullanılmaktadır.

Önerilen jeodezik benzerlik ölçüsü metin tabanlı bağlı veri seti Wikipedia XML Corpus üzerinde denenmektedir. Wikipedia bağ çizgesi üzerinden hesaplanan kümelenme katsayısı değerleri eğrilik değerleri olarak kabul edilmekte ve metinden hesaplanan Cosine benzerlik ölçüsü değerleri ile çember üzerindeki jeodezik formülü kullanılarak birleştirilmektedir. Deneysel çalışma kümelenme

bağlamında gerçekleştirilmiş ve sonuçlar kümelenme metrikleri üzerinden yorumlanmıştır.

Çalışma, bilgi modellemesinde manifold fikrini ortaya koyması ve bu fikrin geçerliliğini manifold fikrine dayanan bir semantik analiz yöntemi ve uygulaması sunarak test etmesi açısından özgündür. Bu niteliği ile gelecekte yürütülecek çalışmalara öncülük etmektedir.

Anahtar sözcükler: Bilgi modeli, semantik analiz, manifold, eğrilik, jeodezik uzunluk, benzerlik ölçüsü.

ABSTRACT

ANALYSIS AND MODELING OF SEMANTIC INFORMATION

TEKİR, Selma

Ph. D. in Computer Eng.
Supervisor: Prof. Dr. Şaban EREN
October 2010, 94 pages

Information has a broad meaning. It can be interpreted as raw material that is directly acquired via observations. At a higher level, information is the results and opinions that are reached directly and/or indirectly through data processing and analysis. In fact, the level of analysis determines the information abstraction level. The semantic analysis can be defined as the methods by which the high-level information characteristics can be addressed in terms of the low-level, directly measurable information attributes. The semantic analysis is a suite of methods, there exists no unique semantic solution.

In this thesis, in order to represent semantic information, the information modeling requirements are considered.

Semantic information can be modeled on a manifold. Manifold is a generalized space than the metric space but through the local properties it provides specialization at the same time. The specialization comes from the position dependence by which curvature metric becomes available in addition to ordinary distance metric.

As part of the work, a manifold based semantic analysis method is proposed in order to support the idea of modeling semantic information on a manifold. The proposed approach is based on the concept of geodesic distance on a manifold. The aim is to utilize geodesic distances for making similarity measurements.

The proposed geodesic similarity measure is tested on the Wikipedia XML Corpus, which is a text-based linked data set. The clustering coefficient values that are computed using the link graph are substituted as curvature values by the assumption that clustering coefficients are rough estimates of curvatures. Then, the curvatures are combined with the Cosine text similarity measure in the

formula of geodesic distance. The experimental work is conducted in the clustering context and clustering metrics are used in order to evaluate the experimental results.

Finally, the work's originality comes from the fact that it proposes the manifold structure for representing information, develops and tests a manifold-based semantic analysis method to justify the information manifold idea. Thus, it opens a new research area and motivates similar approaches for future research.

Keywords: Information model, semantic analysis, manifold, curvature, geodesic distance, similarity measure.

TEŞEKKÜR

Öncelikle tez danışmanım Prof. Dr. Şaban Eren'e süreci benim için kolaylaştırmaya çalıştığı ve bana sonuna kadar güvendiği için teşekkür ederim.

Ortak tez danışmanım Doç. Dr. Ahmet Koltuksuz bilime bakış açısı ve vizyonu ile çalışmayı üst noktalara çekmek için büyük gayret gösterdi, kendisine emekleri için teşekkür ederim.

Türkiye Bilimsel ve Teknik Araştırma Kurumu'na (TÜBİTAK) verdiği yurt içi doktora bursu için teşekkür ederim.

ACM-Mentornet işbirliği sayesinde koçluğumu yapan Dr. Jason Williams gerek resmi koçluk süresi içerisinde gerekse sonrasında fikir ve önerileri ile bana çok yardımcı oldu, yoğun iş temposu içerisinde bana her zaman öncelik verdi. Kendisine teşekkürü bir borç bilirim.

Doktora araştırma ziyareti kapsamında 9 aylık bir süre ile kürsüsünde çalıştığım Prof. Dr. Daniel Keim özel bir teşekkürü fazlasıyla hakediyor. Her zaman, her koşuldaki yapıcı tavrı, iyimser gerçekçiliği, nezaketi ve bilgisi ile benim için çok iyi bir örnek oldu.

Araştırma ziyaretim sırasında idari işlerin yönetimini ele alan, bana öncelikli davranan Dr. Johannes Dingler'e de teşekkürü bir borç bilirim.

Araştırma ziyareti boyunca aynı ofisi paylaştığım Yük. Müh. Slava Kisilevich fikir ve önerileri ile çalışmaya oldukça katkı sağladı, kendisine teşekkür ederim.

Benimle aynı dönemde doktorasını tamamlayan Dr. Daniela Oelke akademik olgunluğu, hemen her konuda çözüm üretme çabası ve dostluğuyla özel bir teşekküre fazlasıyla layık.

Kodlama kısmında karşılaşılan bir problemi aşmada Yük. Müh. Fabian Fischer yardımını esirgemeyip çok kısa bir süre içerisinde çözüm üretmemizi sağladı, desteği için çok teşekkür ederim.

Prof. Dr. Oktay Pashaev tez çalışmasında yapılanları büyük bir dikkatle dinleyip çok değerli fikir ve önerilerde bulundu. Svetlana Pashaeva dostluğuyla hep yanımda yer aldı. Kendilerine teşekkür ederim.

Yrd. Doç. Dr. Enver Tatlıcıoğlu'na da süreç boyunca verdiği destek için teşekkür ederim.

Hem akademik alanda hem de insani olarak güven duyduğum Yük. Müh. Mutlu Beyazıt'ın varlığı da oldukça önemliydi. Birlikte aldığımız modern geometri dersi ve düzenli fikir alışverişlerimiz günden güne bizi geliştirdi ve zorlukları aşma konusunda teşvik edici oldu.

Yrd. Doç. Dr. Serap Atay süreç boyunca hep yanımda yer aldı, dostluğu ve desteği benim için büyük bir güç oldu. Kendisi özel bir teşekküre fazlasıyla layık.

Ayrıca İYTE Bilgisayar Mühendisliği Bölümü'ndeki Hocalarıma ve iş arkadaşlarıma içinde bulunduğumuz ve çalıştığımız güzel ortam için teşekkür ederim.

Son olarak aileme, verdikleri emek ve bana olan sonsuz güvenleri için çok teşekkür ederim.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (devam)

Sayfa

ŞEKİLLER DİZİNİ

Şekil Sayfa

ÇİZELGELER DİZİNİ

Çizelge Sayfa

ALGORİTMALAR DİZİNİ

Algoritma Sayfa

1. GİRİŞ

1.1 Bilgi ve İlişkin Tanımlar

Bilgi her dönemde en önemli güçtür. Özellikle bugün, bilginin asıl güç simgesi olarak kabul edilmesi, yaşadığımız çağa bilgi çağı denmesi, gücü dünyayı yönetme boyutunda algılama geleneğidir. Oysa eski çağlarda güç, kişinin kendi üzerindeki hâkimiyeti olarak algılanmıştır (Williams, 2001).

Bilgi, bilimin ana malzemesi ve ana sonucudur. İnsan algısı ile içiçe geçmiş durumdadır. Bu ikili ilişkide insan yönüne odaklanılırsa, bilginin öznelliğinden bahsedilir; bilgi, insan algısından bağımsız bir şekilde ifade edilmeye çalışıldığında ise nesnel bakış açısı devreye girer. Bilgi felsefesi öznel ile nesnel olan arasındaki sınırı çizmeye çalışır ve daha çok öznel nitelikler üzerine yoğunlaşır.

Kurallara uygun (formal) modelleme yaklaşımları bilginin nesnel olduğunu kabul eder. Öznellik daha yüksek bir soyutlama seviyesinde incelenmelidir ve henüz bu seviyede geçerli olan matematiksel yöntemler mevcut değildir.

Bilgi-insan ilişkisinde geleneksel olarak bilgiden etkilenen, bilgi ile şekillenen insandır. İnsan gözlem yoluyla bilgiyi elde eder. Kuantum teorisi gözlemcinin gözleneni etkileyebileceği varsayımına dayanır ve bilgiye farklı bir boyut kazandırır.

Bilgi belli bir hiyerarşi dâhilinde incelenmektedir. Hiyerarşinin her bir katmanı farklı bir soyutlama seviyesine karşılık gelmektedir. Aşağıda, bilginin her bir soyutlama seviyesindeki karşılığı ve ilgili tanım verilmektedir:

- Veri (Data): Nitel ve nicel gözlemler yoluyla elde edilen unsur.

- Bilgi (Information): Belli bir yapıda bir araya getirilmiş, düzenlenmiş veriler.

- İrfan (Knowledge): Bilginin, bilgi parçacıklarının arasındaki semantik kuralların anlaşılması sonucunda ulaşılan üst düzey bilgi.

- Hikmet (Wisdom): İrfan (knowledge) sahibi kişilerin bilgiler arasındaki semantik kuralları sezmeleri ile sahip oldukları bilgiyi farklı alanlara uygulama yetenekleri (Waltz, 1998).

Bilginin değeri işe yarar, kullanılabilir olması ile ölçülür. Kullanılabilir bilgi üç temel özelliğe sahiptir:

- konu ile ilgili olma (aynı bağlamda olma),
- doğruluk,
- zamanlı olma.

Bilgi felsefesi kapsamında bilgiyi bilgi yapan temel nitelik, doğruluktur. Bilgi, doğruluğu ve bu doğruluğun yeterli dayanaklarla gösterilmesi (objektif yöntemler kullanılarak) gerekçesi ile inançtan ayrılmaktadır (Williams, 2001).

Bilgi modeli bilginin ölçülmesini sağlayan bir çerçeve sunar. Model kapsamında bilgiye ait doğrudan ölçülebilecek özellikler saptanır. Doğrudan ölçülebilecek düşük seviyedeki bütün özellikler işlenerek yüksek seviyedeki bilginin temel niteliklerine erişilir. Bu noktadaki temel kısıtlama insanın ve oluşturduğu yapıların bilginin özelliklerinin çok küçük bir altkümesini algılama ve ölçme yeteneğinde olmasıdır. Dolayısıyla en genel hali ile bir bilgi modeli, bilgi nesnesinin idrak edilen özellikleri üzerinden temsil edilmesi ve işlenmesidir (Muller, 2007).

Bilgi, üç seviyede işlenebilir (Daconta, 2003):

- Sözdizimsel (syntactic).

Formatın ve karakter sırasının önemli olduğu, karakterlerden belirteçlerin (token) oluşturulduğu, belirteç düzeyinde çözümlemenin gerçekleştirildiği seviyedir. Makineye özgüdür, tamamen otomatiktir.

- Anlamsal (semantic).

Bilgi belirteçlerinin (token) kurallar yardımı ile biraraya getirildiği ve değerlendirildiği katmandır. Kuralların karmaşıklık derecesine, niteliğine bağlı

olmak üzere kendi içerisinde seviyelendirilir, yarı-otomatiktir. İnsanın makineye baskın olduğu seviyedir.

- İşlevsel (pragmatic).

Bilgi parçacıklarını birleştiren kuralların net bir şekilde tanımlanamadığı, hikmetin geçerli olduğu seviyedir. Niyeti algılamayı gerektirir. İnsana özgüdür, henüz bu seviyede çalışacak bir bilgi modeli tasarlanamamıştır.

Varolan bilgi modelleri sözdizimsel seviye ile kısmen semantik seviyeyi kapsamaktadır. Bu tez çalışması ile semantik seviyeye karşılık gelen bir bilgi modelinin altyapısı ve nitelikleri hakkında düşünceler ortaya konmakta ve bu modelin bileşeni olarak kullanılabilecek bir semantik analiz yöntemi önerilmektedir.

1.2 Analiz

Gücün yeni simgesi bilgi, bugünkü algılanış şekli doğrultusunda küresel ölçekte yarattığı etkilerle birey ve kurum tanımlarını değiştirmeye başlamıştır.

Günümüzde bireylerin ve kurumların mutlu ve başarılı bir şekilde varlıklarını sürdürebilmeleri sahip oldukları bilgi sisteminin etkili bir şekilde işleyişi ile mümkündür. Bu etkili işleyişin en önemli göstergesi isabetli alınan kararlardır. Bir başka ifade ile bilgi sistemleri karar verme süreçlerini destekleyici bir görevi yerine getirir.

Bilgi sistemleri döngüsel bir işleyişe sahiptir. Bilgi işleyiş döngüsü aşağıdaki gibi olmak üzere iki farklı bakış açısı ile yorumlanmaktadır: Klasik bilgi elde etme (Information Retrieval-IR) sistemleri ve istihbarat analizi.

Bilgi elde etme, büyük derlemler halinde tutulan dokümanlardan bilgi gereksinimini karşılayacak, genellikle yapılandırılmamış materyalin bulunmasıdır (Manning et al., 2008).

İstihbarat ise karar verenlerin, gerekli bilgi ile desteklenmesini sağlayan temel bir işlevdir. Karar verenlerin taleplerini karşılamak üzere istihbarat analistleri tarafından oluşturulur. Geleneksel istihbarat döngüsü; istihbarat problemine ait gereksinimlerin belirlenmesi, kaynak tahsisinin yapılması, veri

toplanması, toplanan verinin değerlendirilmesi (analizi) ve karar verilmesi aşamalarından oluşur (U.S. Congress, 1996).

İstihbarat döngüsünün en önemli aşaması, insanın ön plana çıktığı, kararın oluştuğu analiz aşamasıdır. İstihbarat analizi politika hedefi dikkate alınarak bir istihbarat probleminin çözümlenmesi, kısaca doğasının ortaya çıkarılması çabasıdır. Analiz, istihbarat analistleri tarafından yapılır, elde edilmiş veriye değer katan yorumlar içerir.

İstihbarat analizi karar verenin açıkça ortaya koyduğu istihbarat problemini aydınlatmak üzere analistler tarafından yürütülür. Bilgi elde etme sistemlerinde ise nesnellik ön plandadır; bilgi arayışında olan kişi ve öznel nitelikler ikincil durumdadır.

İstihbarat analizi sistemleri analistin işini kolaylaştırmak amacıyla analiste kişiselleştirilebilir analitik ve işbirliği araçları sunar (Kamarck, 2005). Analistin süreçte merkezi bir rolü vardır.

İstihbarat döngüsünün araştırma-keşfetme safhası istihbarat probleminin doğasının keşfedildiği aşamadır. Analizin bileşeni olan içgörüler (insights) bu araştırma aşamasında gelişir. İstihbarat analizi bağlamında, bilgi elde etme yorumundan farklı olarak tek tek sorgular yerine birbirini takip eden sorguların arasına yerleşmiş fikirlerin sorgularla birlikte karakterize ettiği keşif niteliğinde bir arama eylemi sözkonusudur (Gersh et al., 2006).

Arama sonuçlarından ziyade, bizzat arama sürecinin kendisi gerekli içgörüyü verir. Dolayısıyla istihbarat analizi yinelemeli (iterative) arama süreci üzerine odaklanır. İstihbarat analizinin odaklandığı bu konular, bilgi elde etme sistemlerinde de popüler hale gelmektedir.

İstihbarat analizi ve bilgi elde etme sistemleri arasındaki karşılaştırmanın bir benzeri Swanson (1988) tarafından istihbarat analisti ve bilim insanı arasında yapılmaktadır. Sözkonusu çalışmada, istihbarat analisti ve bilim insanı, kaydedilmiş bilgiyi kullanımları yönünden değerlendirilmektedir.

Bu değerlendirme öncesinde bilimin geçen yüzyılda bilgi ile ilişkisini belirtmekte fayda vardır. Sanayi toplumuyla birlikte bilgi bilim yolu ile yaratılmaya başlanmış ve örgün eğitim-öğretim ile aktarılmıştır. Bilgi basılı

dokümanlarda tutulmuştur. Bilgi toplumunda basılı dokümanlara ek olarak elektronik materyaller hatta yazılımlar eğitim-öğretim materyali olarak karşımıza çıkmaktadır; internet eğitim-öğretim dünyasına damgasını vurmuştur. Özetle günümüzde bilim insanının hizmetine sunulmuş çok sayıda kaynak ve servis bulunmaktadır (Cellary, 2003).

Bilim insanlarındaki genel eğilim; yeni çalışmalar yapmaya, yeni ürünler yaratmaya öncelik verirken, daha önce yapılmış çalışmaları okumayı, varolan çalışmaları derleyici, açıklayıcı raporlar yazmayı kısa sürede tamamlanması gereken zorunlu işler olarak görme yönündedir. Bilim insanı kendi uzmanlık alanındaki gelişmeleri takip etmenin yeterli olacağına inanmaktadır.

İstihbarat analistiyse; bilim insanının aksine, bilgi ile daha yakın bir iletişim içerisindedir. Farklı bilgileri tek başlarına büyük anlam taşımasalar dahi anlamlı bir bütünün parçaları olma potansiyellerini dikkate alarak inceler. Bir konu ile ilgili dokümanlar aramak yerine senaryolar, örüntüler (pattern) arandığından ve bu örüntülere ait parçalar bulunmaya çalışıldığından, bu görüş bilgi elde etme sistemlerinin yetersiz bir metafor olduğunu destekler. İdeal bir bilgi sistemi, gereksinim duyulan bilgiyi zamanında almalı, gerekiyorsa arşivlemeli ya da erişim kanallarını kaydetmeli, işlemeli, değerlendirmeli (analiz) ve değerlendirme sonucu olan kararı diğer bilgi sistemleri ile etkileşiminde kullanmalıdır. Varolan - neredeyse sınırsız- bilgi ve iletişim olanakları, içeriye doğru bilgi akışını yoğunlaştırmakta ve işleme kapasitesinin sınırlarını zorlamaktadır. Giriş kanallarını yönetmek dahi başlı başına bir sorun haline gelmektedir. Seçilmiş iletişim kanalları üzerinden gelen bilgiyi değerlendirmek ise sürecin en zor safhasıdır; çünkü bilgi hem miktar olarak çok büyüktür, farklı formlardadır ve hem de doğrulanmış değildir. Bütün bu zorluklar bireyin ya da kurumun bilgi işleme sürecinde yazılım ile desteklenmesi yoluyla hafifletilebilir/aşılabilir. Bir başka deyişle; yazılım bilgi sistemlerinin vazgeçilmez bir bileşenidir ve bilginin etkili bir şekilde kullanımını kolaylaştırır.

Bilgi işleme döngüsünün en önemli aşaması, insanın ön plana çıktığı, kararın oluşturulduğu analiz aşamasıdır ve kendisinden önce gelen aşamaları yeniden tanımlar. Analiz diğer safhalarla bütünleşik bir nitelik taşımaktadır. Bununla birlikte genel geçer bir analiz yöntemi mevcut değildir. Hedefe, elde edilen verinin yapısına bağlı olarak çok çeşitli analiz yöntemleri vardır. Bu yöntemlerin tek tek ya da birlikte kullanılması ve yorumlanması analizi meydana getirir. Analiz safhasını besleyecek olan bilginin uygulanacak yöntem(ler)e uygun

formda olması gerekir. Çoğu zaman verinin uygun formda olup olmayışı analizin başarısını belirler.

Bilgiyi işleyip analiz etmede kullanılabilecek birçok yazılım/yazılım aracı bulunmaktadır. Burada göz önünde bulundurulması gereken nokta, analiz yapma yeteneğinin hâlâ insana özgü bir yetenek olduğu, yazılımın bu anlamda insanın yerini şu an için al(a)madığı ve ancak kolaylaştırıcı, destekleyici, tamamlayıcı olarak sürece dâhil olduğudur. Bir başka önemli husus, ilgili yazılımların bilginin doğası, nitelikleri gözönünde bulundurularak tasarlanmasıdır. Ayrıca standart bir analiz yöntemi yoktur, çok çeşitli analiz yöntemlerinin oluşturduğu bir bütün sözkonusudur. Her yöntem girdisine uygun şekilde tanımlanmalı ve semantik seviyede çalışan matematiksel fonksiyonlarla zenginleştirilmelidir.

Bu tez kapsamında; analiz öncesindeki bilgi ile "information" kastedilmekte, analiz sonrasında sözcüğün irfan karşılığı kullanılmaktadır.

1.3 Semantik Analiz

Semantik analiz; olguların (phenomena) arkasında yatan, onları açıklamaya yarayan kavramların ortaya konmasıdır. Örnek bir olgu olarak gerçek sosyal ağların büyümesi ele alınabilir. Sosyal ağ bir sosyal etkileşim ve ilişkiler sistemidir, sosyal bağlarla birbirine bağlı insan topluluğudur (Oxford English). Gerçek sosyal ağlarda çok sayıda bağlantıya sahip olan kişilerin yeni bağlantılar kurma olasılıkları yüksektir. Hatta kişinin çevresi oldukça geniş ise yeterince uzun bir sürede (sonsuza yaklaşırken) toplumun tüm bireylerini tanıması mümkün hale gelmektedir. Bununla birlikte bu kadar çok sayıda tanıdığı olan kişilere rastlanma olasılığı bağlantı sayısının karesi ile ters orantılıdır. Bu iletişim davranışı matematiksel olarak power-law dağılışı ile açıklanabilir. Dolayısıyla, sosyal ağlarda büyüme olgusunu modellemeyi sağlayan power-law dağılışı, semantik analiz yöntemlerinden biridir (Adamic et al., 2001).

Sosyal ağlardaki bu iletişim davranışı bir mühendislik ürünü olan web ve benzeri ortamlarda da gözlenmektedir. Web'i oluşturan dokümanların bağ sayıları power-law dağılışı gösterir. Web üzerindeki bir dokümanın diğer k sayıdaki dokümana bağlanma olasılığı $P(k) \sim k^{-\gamma}$ formülü ile açıklanabilir. Web iç-bağ (in-link) sayıları dikkate alındığında γ değeri 2.1 olarak hesaplanmıştır (Barabási et al., 2000). Gerek doğada gerekse insan tarafından yaratılmış sanal bir ortamda gözlenen bu ortak davranış, bilginin benzer nitelikleri taşıması durumunda aynı

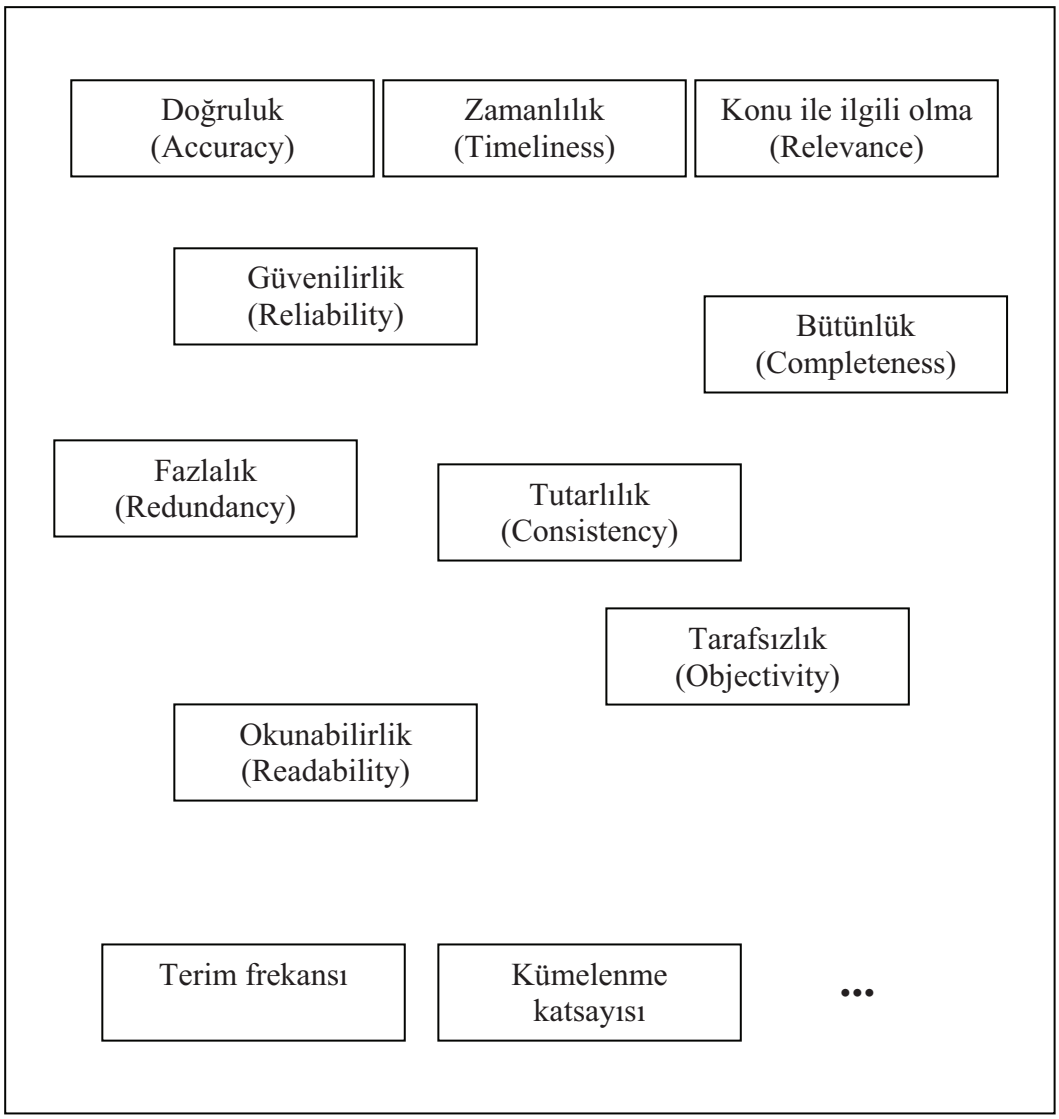
semantik analiz yönteminin farklı ortamlarda geçerli olabileceği sonucunu ortaya koyar. Temel benzerlik noktası her iki ortamda da k bağlantı sayısının hesaplanabilir olmasıdır. Basit sayma işlemi ile hesaplanabilen bu tip bilgi özellikleri düşük seviyedeki özellikler (low-level features) olarak adlandırılır. Semantik analiz yöntemleri, benzer düşük-seviyedeki özelliklerin (low-level features) hesaplanabileceği bütün veri setleri üzerinde geçerli olma potansiyeli taşır. Hedef, kullanılan semantik analiz yöntemi ile düşük seviyedeki özelliklerden yola çıkarak yüksek seviyedeki özellikleri (high-level features) keşfetmektir.

Semantik analiz bilgisayar bilimlerinde farklı bağlamlarda farklı anlamlara karşılık gelir. Örneğin makine öğreniminde (machine learning) bir külliyatın (corpus) semantik analizi o külliyatı oluşturan doküman kümesini temsil eden kavramları çıkaracak yapıların oluşturulması işidir. Semantik web alanında, belli bir iş alanını (domain) temsil eden ontolojinin yaratılması bir çeşit semantik analizdir. Bu tez kapsamında, semantik analiz makine öğrenimindeki anlamına benzer şekilde bir doküman setine ait düşük seviyedeki sözdizimsel özelliklerin birleştirilerek bazı yüksek seviyede özelliklerin keşfedilmesi anlamında kullanılmaktadır. Bir külliyat için düşük seviyedeki özelliklere örnek olarak külliyatın içerdiği toplam sözcük sayısı, bulunan her sözcüğün külliyattaki ve o külliyatı oluşturan dokümanlardaki frekans değeri, dokümanlar arasındaki bağ (link) sayıları verilebilir. Yüksek seviyedeki özellikler ise bir dokümanın okunup anlaşılabilirliği (readability-understandability), külliyatın güvenilirliği (reliability), külliyatta adı geçen bir kişi üzerine oluşan izlenimler (opinions, sentiments) ya da külliyat içerisindeki yetkili, güvenilir (authoritative) kaynakların tespit edilmesi olarak örneklenebilir.

Semantik seviyede çalışacak bir bilgi modeli, yüksek seviyedeki özellikleri keşfetmemizi sağlayan farklı semantik analiz yöntemlerinin birarada değerlendirilmesi sonucunda mümkün olabilir. Semantik analiz yöntemleri geliştirildikçe bu yöntemlerin ortak özellikleri ve farklılıkları görülerek hepsini temsil eden bir gösterim oluşturulabilir. Bir başka ifade ile bir üst soyutlama düzeyine çıkılabilir, ancak bu hedef şu an için uzak bir hedeftir.

Semantik bir bilgi modeli bakış açısıyla bilgi ve nitelikleri değerlendirildiğinde, bilgi nitelikleri bir hiyerarşide gösterilebilir. Hiyerarşinin en üst seviyesinde bilginin temel nitelikleri olan konu ile ilgili olma, doğruluk ve zamanlılık yer alır. En alt katmanlarda ise doğrudan ölçülebilen düşük seviyeli

özellikler görülür. Şekil 1.1'de karmaşıklıkları dikkate alınarak, bilginin nitelikleri bir hiyerarşide gösterilmektedir:



Şekil 1.1 Bilgi Nitelikleri Hiyerarşisi.

Adı geçen bilgi niteliklerine ilişkin tanımlar aşağıda verilmektedir (Batini and Scannapieco, 2006):

Doğruluk: Bir değerin temsil ettiği kavramı doğru bir şekilde karşılayan bir diğer değere yakınlığıdır.

Zamanlılık: Bilginin varolan amaç için ne kadar güncel olduğunun göstergesidir.

Konu ile ilgili olma: Verilen bir bilgi nesnesinin aranılan bilgiye uygunluğunun ölçüsüdür.

Güvenilirlik: Bir kaynağın doğru bilgiyi iletip iletmediğinin ölçüsüdür. Böylece ilgili kaynaktan elde edilen bilgiye güvenilebilir.

Bütünlük: Verinin belirlenen amaç için yeterli genişlikte, derinlikte ve kapsamda olma derecesidir.

Fazlalık: Bilginin gereksinimin ötesinde tedarik edilme derecesidir.

Tutarlılık: Biraraya getirilen bilgi öğelerinin birbiri ile çelişmemesidir.

Tarafsızlık: Bilginin elde edildiği kaynağın yansız olma ölçüsüdür.

Okunabilirlik: Bilginin temsil ettiği kavramı planlanan kullanım çerçevesinde açık ve net bir şekilde ifade etme ölçüsüdür.

Terim frekansı: Bir terimin bir dokümanda tekrar etme sayısıdır.

Kümelenme katsayısı: Bir çizge üzerinde yer alan bir düğümün yerel bağlantısallık derecesinin göstergesidir. İlgili düğümün komşuluğundaki kenar sayısı baz alınarak hesaplanır.

Bilgiyi modellemek üzere bu tezde önerilen yaklaşım; aşağıdan yukarıya (bottom-up) olup basitten karmaşığa doğru özellik hiyerarşisinin oluşturulmasıdır. Temel hedef alt katmanlarda yer alan özelliklerin birleştirilmesi suretiyle en üst katmandaki özelliklere varmaktır. Bilginin en soyut üç niteliğine erişildiği takdirde modelleme amacına ulaşmış olur.

Semantik bilgi modelinin yapıtaşı olacak semantik analiz yöntemleri geliştirilirken dikkat edilmesi gereken bir diğer husus da her semantik analiz yöntemi ile bir bilgi niteliğinin birebir eşlenemeyeceğidir. Bazı yöntemler birden fazla niteliği bünyesinde barındırabilir.

Semantik analiz yöntemleri olguları (phenomena) karşılamaya yönelik olduğundan, zaman kavramı bir ölçüde semantik analiz kavramı içerisinde değerlendirilebilir. Olgular olaylardan farklı olarak anlık değildir, süreklilik

niteliği taşır. Dolayısıyla semantik bilgi modelinin süreklilik niteliğini içsel olarak barındıran semantik analiz yöntemleri, zamanı mutlak olarak değerlendiren yapıtaşlarından meydana gelmektedir.

Sözdizimsel-semantik seviyede bir modelleme sağlayan günümüz bilgi modelleri metrik uzaylar üzerine yerleşmiştir ve uygulamaları mevcuttur. Bilginin kuantum mekaniği temel alınarak modellenmesi gerektiğini ortaya koyan görüş, karmaşık sayı metrik uzayını varsayılan olarak kabul eder, ancak bilinen bir uygulaması henüz yoktur (Rijsbergen, 2004).

Bu tez çalışmasının temel fikri, bilgiyi bir manifold üzerinde modellemektir. Manifold hem Öklidyen hem de Öklidyen olmayan geometriyi birlikte temsil eden, gerek yerel gerekse global özelliklerin desteklenip yerel özellikler arasında dönüşümlerin yapılabildiği bir ortamdır. Eğrilik metriği geçerlidir. Değişken eğrilik değerleri dolayısıyla farklı uzay tanımları bileşen olarak yerini alır. Çalışmada bilginin manifold üzerinde temsil edilmesi fikri bir ölçüde doğrulanmaya çalışılmaktadır. Bu amaçla kullanılan yöntem manifoldlar üzerinde geçerli olan jeodezik uzunlukların varolan modellere bir katkı sağlayıp sağlamadığını göstermektir. Jeodezik uzunluklar varolan uzunluk kavramına katkı sağladığı takdirde hem bilginin manifold üzerinde modellenmesi fikri desteklenmiş olacak hem de spesifik bir semantik analiz yöntemi yolu ile yüksek seviyedeki bir bilgi özelliği daha karşılanacaktır.

Dolayısıyla; bu çalışmada, iki yönlü bir hedef mevcuttur:

- Bilginin üzerine yerleştiği matematiksel yapı hakkında ipucu elde etmek.

- Semantik bir bilgi modeli fikrini, bileşenini oluşturacak bir semantik analiz yöntemi ile desteklemek.

İkinci hedef; semantik bilgi modelini, semantik analiz yöntemlerinin bir bütünü, yüksek seviyedeki özelliklerin iç içe kullanılması yolu ile ulaşılmaya çalışılan üç temel bilgi niteliğinin temsil edilmesi olarak gören bakış açısının ürünüdür.

Tez kapsamında önerilen semantik analiz yönteminin özünü oluşturan fikir ise, Öklidyen olmayan geometriye dayanmaktadır. Öklid geometrisi, içinde bulunulan uzayın düz (eğriliğin 0) olduğu varsayımıyla Öklidyen olmayan

geometrinin özel bir hali gibi düşünülebilir. Öklidyen olmayan geometride uzay eğimlidir ve bu uzayda geçerli olan metrikler varolan eğrilik değerlerini dikkate almaktadır. Örneğin; jeodezikler (iki nokta arasındaki en kısa eğriler) eğrilik tabanlı hesaplanır. Bu tezdeki ana hipotez; jeodezik uzunlukların anlamsal yakınlıkları (semantic relatedness) ölçmek üzere kullanılabileceğidir. Uygulama, bağlı (linked), metin tabanlı (text based) bir veri seti üzerinde gerçekleştirilmektedir. Veri seti, Wikipedia XML Corpus (Denoyer and Gallinari, 2006) olarak adlandırılan külliyatın ingilizce Wikipedia altkümesinden oluşmaktadır. Veri seti üzerinde metin tabanlı analiz yapılarak terim-doküman matrisi elde edilmiş, dokümanlar arasındaki bağlar kullanılarak da dokümanların iç-bağ (in-link) ve dış-bağ (out-link) dağılışları oluşturulmuştur. Bu analizler sözdizimsel seviyeye karşılık gelmektedir. Semantik seviyeye çıkmak için daha yüksek seviyedeki yapıları keşfetmek gerekmektedir. Uygulanan yaklaşımda bağ çizgesi üzerinden hesaplanan kümelenme katsayıları (clustering coefficients) ile terim-doküman matrisinden elde edilen Cosine benzerlik ölçüsü değerleri, jeodezik uzunluk formülünde birleştirilmiştir. Buradaki temel varsayım, kümelenme katsayısı değerlerinin eğrilik değerlerini tahminlemeleridir (Lou, 2009). Oluşan jeodezik uzunluk metriği veri seti üzerinde anlamsal yakınlık (semantic relatedness) kavramını açıklamak için kullanılmıştır.

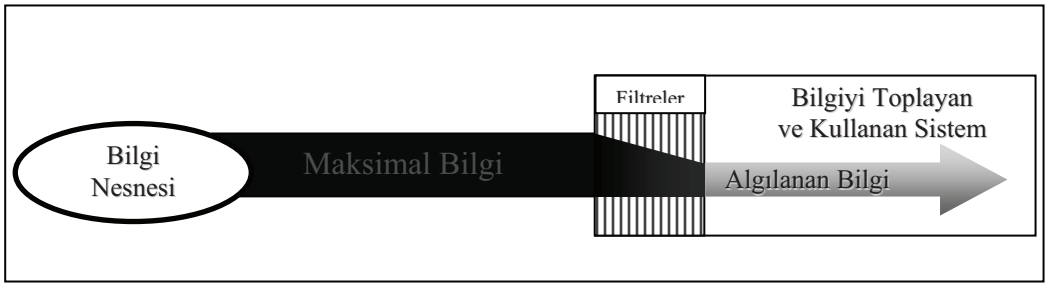
Tezin bir sonraki bölümünde, geleneksel bilgi modellerine ait altyapı incelenmektedir. Bu bölümde cebirsel yaklaşımlara ağırlık verilmiş olup, özellikle Öklidyen olmayan modeller ve yöntemler irdelenmiştir.

2. BİLGİ MODELLERİ ALTYAPISI

2.1 Giriş

Modeller; kavramların sayısal olarak temsil edilmesini ve oluşturulan sayısal karşılıklar üzerinden yapılan hesaplamalarla, ilgili kavramlar arasında ilişki kurulmasını sağlar. Kavramlara sayısal karşılıklar bulma işlemi bir koordinat sisteminin kullanımını gerektirir. Kavramlar arasındaki ilişkiyi ölçmek için yapılacak hesaplamalar ise koordinat sistemlerinden bağımsız gösterimlere (metrikler) gereksinim duyar.

Bilginin modellenmesi sözkonusu olduğunda bilgiye sayısal bir karşılık bulma ihtiyacı doğar. Bu ihtiyacı karşılamak üzere bilgiye ait doğrudan ölçülebilecek özellikler bulunmalıdır. Doğrudan ölçülebilecek düşük seviyedeki bütün özellikler işlenerek, yüksek seviyedeki bilginin temel niteliklerine erişilir. Bu noktadaki temel kısıtlama insanın ve oluşturduğu yapıların bilginin özelliklerinin çok küçük bir altkümesini algılama ve ölçme yeteneğinde olmasıdır. Dolayısıyla en genel hali ile bir bilgi modeli bilgi nesnesinin idrak edilen özellikleri üzerinden temsil edilmesi ve işlenmesidir (Şekil 2.1) (Muller, 2007).



Şekil 2.1 Bilgi Modeli (Muller'den 2007).

Bilginin modellenmesi hususunda bir başka temel gereksinim, bilgiyi içerdiği kavramlara ayrıştırmak ya da bilgiden bilgi parçacıkları oluşturmaktır çünkü ölçme işlemi birim cinsinden yapılır ve bilgiyi birimler ile ifade etmenin yolu da budur. Bilgi parçacıklarını belirledikten sonra onlara sayısal karşılıklar bulmak, bulunan sayısal değerler üzerinde yapılan işlemlerin bilgi parçacıkları ile ilişkisini kurmak gerekir. Karşılaşılan temel problemlerden ilki modelin yapıtaşı olacak bilgi parçacıklarının büyüklüğünü (granularity) belirlemektir. Sözcükler genellikle kavramları ifade ettiğinden sözcük tabanlı yaklaşımlar yaygındır. Oysa

problem bilginin doğrulanması olarak ortaya konduğunda, bir yargıyı ortaya koyan önermeler, bileşen olarak ön plana çıkar.

Bir bilgi modeli oluşturulurken görülen bir başka önemli problem, farklı ortamlardan gelen farklı tipteki verilerdir. Metin, görüntü, ses, video bu duruma örnek olarak verilebilir. Öte yandan, her tip veriyi kapsayacak genel geçer bir çerçeve (framework) tanımlama ihtiyacı da vardır. Bilgi parçacıkları da genellikle bilginin tipine göre belirlenmektedir. Örneğin; görüntülerin yapıtaşları, görüntü öğeleridir (pixel).

Ele alınması gereken bir diğer faktör, bilgiye ait üst bilgilerin kullanılacak modelde temsil edilmesi ve uygun şekilde değerlendirmeye katılmasıdır. Bilginin elde edildiği kaynak tipik bir üst bilgidir.

Bilgi parçacıkları en genel haliyle bir kümenin elemanları olarak temsil edilir. Her bilgi parçacığı bir nesnedir. Nesneler üzerindeki ilişkilendirmeler metrikler yoluyla gerçekleştirilir. Bu nispeten soyut kurulum, metrik uzaylara karşılık gelir ve varolan bilgi modelleri geometrik yapı olarak metrik uzaylar üzerinde geliştirilmiştir. Metrik uzaylarda temsil edilen bilgi öğelerinin bulunduğu kümeye X denirse; X kartezyen çarpım kümesinden reel sayılar kümesine tanımlanmış bir d fonksiyonu, bilgi karşılaştırması için kullanılabilir (Chávez et al., 2001):

$$d : \mathrm{X} \times \mathrm{X} \to \Re \qquad (2.1)$$

Bilgi öğeleri arasındaki uzaklığı ölçmesi beklenen bu fonksiyonun tutarlı karşılaştırmalar sağlaması için metrik özellikleri taşıması gerekmektedir. Bu özellikler aşağıda listelenmiştir (Chávez et al., 2001; Searcóid, 2007):

1. $\forall x, y \in X, d(x, y) \geq 0$ pozitiflik,

2. $\forall x, y \in X, d(x, y) = d(y, x)$ değişme,

3. $\forall x \in X, d(x, x) = 0$ yansıma,

ve çoğu durumda

4. $\forall x, y \in X, d(x, y) > 0$ katı pozitiflik,

5. $\forall x, y, z \in X, d(x, y) \leq d(x, z) + d(z, y)$ üçgen eşitsizliği.

2.2 Geleneksel Bilgi Modelleri

2.2.1 Temel kavramlar

Bilginin modellenmesi, pragmatik bir bakış açısı ile aranan bilgiye zamanında erişimi sağlamaktır. Bu tanımda temel varsayım aranılan bilginin aramanın yapılacağı veritabanında bulunduğudur ve hedef o veritabanından istenilen bilgiyi çekmektir. Bu pragmatik çerçeve içerisinde tanımlanan modellere bilgi elde etme (Information Retrieval-IR) modelleri denilmektedir. Bilgi elde etme resmi tanımı ise; her türlü veritabanı içerisinde yer alan dokümanlar arasında metin, ses, görüntü vb. tipteki bilgilerin, dokümanları tanımlayan üst bilgilerin, bizzat dokümanların aranmasını kapsayan sanat ve bilim alanıdır (Webster).

Bir IR sistemi kullanıcının bilgi gereksinimini gözönünde bulundurarak ilgili veritabanında yer alan dokümanlar arasından işe yarayabilecek dokümanları tespit etmeye çalışır. Bu amaç için sistemin üç ana görevi gerçekleştirmesi gerekir:

- Doküman kümesinin kavramsal ilişki yapısını ortaya koymak üzere analiz edilmesi.

- Kullanıcı gereksinimlerinin karakterize edilmesi.

- Bir dokümanın kullanıcı gereksinimlerini karşılayıp karşılamadığı konusunda bir değerlendirme yapılması.

IR sistemlerinde temel problem, kullanıcının bilgi gereksinimini tanımlamaktır. Herhangi bir andaki bilgi gereksinimini içinde bulunulan bağlam belirler. Bu yüzden asıl önemli olan nokta kullanıcının içinde bulunduğu süreci karakterize etmektir (Wong et al., 1987).

Bilgi elde etme modellerinde bilgi nesneleri dokümanlardır ve dokümanların kavramsal yapısını oluşturmak üzere dokümanlara bir dizi dönüşüm işlemi uygulanır. İlk dönüşüm dokümanları yapıtaşı olan daha küçük bilgi parçacıklarına (sözcüklere) ayırmak ve her sözcüğü bir sayı ile ilişkilendirmektir. Sözcükle ilişkilendirilecek sayı genel olarak doküman içerisindeki sözcük frekansıdır. Dokümanların sözcükler cinsinden ifade edildiği bu gösterim, sözcük torbası (bag

of words) olarak isimlendirilir. Bir sonraki dönüşüm sözcük frekanslarının normalize edilmesini sağlar ve dokümanın esas tanımını (description) ortaya koyar. Şekil 2.2'deki örnekte, bir doküman metni, geçirdiği dönüşümler ve her dönüşüm sonrasında oluşan doküman formları gösterilmektedir:

Metin:

Research in the probabilistic theory of information retrieval involves the construction of mathematical models. In this kind of theory construction the assumptions laid down…

Atlanacak kelimelerin (stopword) kaldırılması ve sözcük gövdelerine inilmesi:

Research probabil theory informat retriev involv construct mathemat model kind theory construct assume lay down

Sözcük torbası (bag of words) gösterimi:

(research,1), (probabil,1), (theory,2), (informat,1), (involv,1), (construct,2), (mathemat,1), (model,1), (kind,1), (assum,1), (lay,1), (down,1),

Tanım (description):

(research,0.5), (probabil,0.5), (theory,1), (informat,0.5), (involv,0.5), (construct,1), (mathemat,0.5), (model,0.5), (kind,0.5), (assum,0.5), (lay,0.5), (down,0.5),

Şekil 2.2 Bir örnek üzerinden doküman metni, gösterimi ve tanımı (Fuhr'dan 2009).

Verilen örnekte, sözcük torbası gösteriminde belirtilen frekans değerleri doküman içerisindeki en yüksek terim frekansına bölünerek normalize edilmek suretiyle doküman tanımındaki 0 ile 1 arasında değişen değerlere ulaşılmıştır. Örneğin (probabil,1) ikilisindeki 1 değeri, en fazla yinelenen "theory" sözcüğünün frekansı olan 2'ye bölünerek doküman tanımında (probabil,0.5) olarak yerini almıştır.

Bilgi elde etme sistemlerinde klasik kullanıcı gereksinimi nesnesi sorgudur. Sorgu da doküman ile aynı dönüşümleri geçirir. Sistemin hedefi; kullanıcının sorgusunu alarak, bu sorguyla ilgili dokümanları ilgililik sırasına göre kullanıcıya döndürmektir.

Sistem, sorgu ile ilgili olan dokümanları bulma işlemini doküman ve sorgu nesneleri üzerinde değil ve fakat geçirilen dönüşümler sonrasında doküman ve sorgu tanımları üzerinde gerçekleştirir. Dolayısıyla sistemin başarımı, doküman ve sorgu tanımlarının doküman ve sorgu nesnelerini ne ölçüde temsil ettiği ile doğrudan ilişkilidir.

Bu genel çerçevenin ortaya konmasının ardından bilgi elde etme sistemlerinin tasarımını şekillendiren temel sorunlar aşağıdaki biçimde listelenmektedir (Dubin, 2004):

- Belli bir kümenin iyi tanımlanmış bir tümleyeninin olup olmadığı.

- Sorgu uzayı ile doküman uzayının aynı olup olmadığı; bir başka ifade ile olası sorgu tanımları kümesinin olası doküman tanımları kümesine eş olup olmadığı.

- Doküman ve sorgu tanımlarının yapısallığı ve birbirinden bağımsız olup olmadığı.

- Sorgular üzerindeki sıralama ilişkisinin dokümanlar üzerinde de benzer bir ilişkiyi gerektirip gerektirmediği (daha özelleşmiş bir sorgunun daha az sayıda, özelleşmiş dokümanları getirip getirmeyeceği).

- Sistemin doküman tanımlama dilinden (document description language) farklı bir sınıflandırma dili içerip içermediği, içeriyorsa bu dil kullanılarak oluşturulan kategoriler ile doküman ve sorgular arasında eşleştirme yapacak fonksiyonların var olup olmadığı.

- Tanımlama dili elemanlarının pozitif nitelikler taşıyıp taşımadığı ve değilleme işleminin (negation), diğer özelliklerden bağımsız bir şekilde gösterilip gösterilemeyeceği.

Mevcut teorik altyapı bağlamında değerlendirildiğinde geleneksel bilgi elde etme sistemlerinin, kullanıcı gereksinimini karşılayacak dokümanları eksiksiz ve fazlasız olarak tespit etme yeteneğine sahip olmadığı görülür. Bu gerçek dikkate alınarak, bilgi elde etme sistemlerine derlemdeki dokümanlara kullanıcının işine yarama olasılığına göre azalan bir sıra verecek şekilde bir mekanizma dahil edilmiştir. Kullanıcıya gereksinimlerini karşılama olasılığı en yüksek olan

dokümandan başlayarak giderek azalan olasılık değerleri doğrultusunda bir doküman listesi döndürülür. Olasılık değerlerine göre derecelendirme yapan bu mekanizma, olasılık derecelendirme prensibine (probability ranking principle) (Robertson, 1997) dayanmaktadır. Olasılık derecelendirme prensibi, olasılık tabanlı bilgi elde etme yöntemlerinin temel bakış açısını ortaya koyar. Doküman sorgu tanımları arasındaki karşılaştırma ve değerlendirmeler, bir sorgu tanımı ile yakından ilişkili olma olasılığına göre dokümanları sıralama hedefi doğrultusunda işletilir. Mükemmel bir değerlendirme, sorgu tanımlarından yola çıkarak sorgularla ilişkili dokümanları sıralamayı sağlar. Optimum bir değerlendirme ise sorgu tanımlarından hareketle doküman seviyesi yerine doküman gösterim seviyesinde (Bkz. Şekil 2.2) bir sıralama oluşturur.

Doküman derlemi içerisinden kullanıcının gereksinimine cevap verecek doküman kümesi elemanlarının kesin olarak saptanamadığı bir senaryoda, kullanıcıya değerlendirmek üzere tüm dokümanlar da sunulamayacağından, olasılık derecelendirme prensibine dayalı liste, kullanıcının tanımlayabileceği bir durma noktası sağlaması açısından da faydalıdır. Kullanıcı gerektiğinde tanımladığı sorguyu genişleterek ya da daraltarak kendisine dönen listenin büyüklüğünü ayarlayabilir.

Bilgi elde etme sistemleri iki ölçütle değerlendirilebilir: (i) konu ile ilgili dokümanların derlemden ne kadar başarılı bir şekilde çekildiği ve (ii) ilgili olmayanların ne kadar iyi bir şekilde reddedildiği. Doküman derlemi içerisindeki kullanıcı gereksinimine uygun dokümanların sistem tarafından getirilme oranı, geri çağırma (recall) olarak isimlendirilirken bir sorgu sonucunda çekilen doküman kümesi içerisindeki işe yarar doküman oranı duyarlık (precision) ile ifade edilir (Rijsbergen, 1979):

$$\text{Geri çağırma} = \frac{\left|\{\text{ilgili dokümanlar}\} \cap \{\text{getirilen dokümanlar}\}\right|}{\left|\{\text{ilgili dokümanlar}\}\right|}$$

$$\text{Duyarlık} = \frac{\left|\{\text{ilgili dokümanlar}\} \cap \{\text{getirilen dokümanlar}\}\right|}{\left|\{\text{getirilen dokümanlar}\}\right|}$$

Bilgi elde etme sistemleri bilginin konu ile ilgili olma (relevance) niteliği etrafında geliştirilmiştir. Doğruluk-hassasiyet (accuracy) ve zamanlılık (timeliness) doğrudan hedef alınan nitelikler değildir. Oysaki işe yarar bilgi üç temel bilgi niteliğinin aynı anda karşılandığı bir model üzerinden elde edilebilir.

Bu noktada cevaplanması gereken temel soru; işe yarar bilgiyi modellemek için modelleme sürecinin başından itibaren bu üç temel bilgi niteliğinin hedef olarak ortaya konmasının ve süreç boyunca gerçekleştirilmeye çalışılmasının gerekip gerekmediğidir. Bir bilgi modeline sonradan bu özelliklerin dâhil edilip edilemeyeceği sorgulanmalıdır. Bu sorunun cevabına göre farklı bilgi niteliklerini hedef alan bilgi modellerinin biraraya getirilmesi mümkün olabilir. Bununla birlikte yazılım geliştirme süreçlerinde kalite hedefi olarak belirlenen soyut özelliklerin (güvenlik, güvenilirlik vb.) projenin başlangıcından itibaren gözönüne alınmadığı takdirde süreç sonunda oluşturulan ürünün birer niteliği olamayacakları yaygın bir kanıdır. Dolayısıyla iki alandaki benzerlik dikkate alınarak işe yarar bilginin belirtilen üç niteliğini aynı anda hedef alan yeni bir bilgi modelinin tasarlanmasına ihtiyaç olduğu söylenebilir.

Bilginin tüm niteliklerini aynı anda karşılayan bir bilgi modeline duyulan gereksinim, konu ile ilgili olma niteliğini kapsayan bilgi elde etme modellerinin önemini yadsımaz. Belirtilen bilgi gereksinimine göre “ilgisiz” olarak değerlendirilen iki dokümanın, biraraya getirildiğinde “ilgili” duruma gelebileceği, henüz varolan modeller kapsamına girebilmiş değildir.

2.2.2 İkili bağımsızlık elde etme modeli (Binary independence retrieval model)

Bilgi elde etme modellerindeki temel varsayım, dokümanların ve sorguların indekslenmiş olmasıdır. Bir başka deyişle, modeller, dokümanlar ve sorgular yerine doküman ve sorgu tanımları üzerinden işleme yeteneğine sahiptir. Doküman ve sorgu kümelerini karşılaştırmak üzere uygulanan en temel yaklaşım, sorgu sözcüklerinin dokümanlar içerisinde aranması yani doküman sözlüğü ve sorgu sözcükleri kümesi üzerinde kesişim işleminin uygulanmasıdır. Sorgu birden fazla sözcük içeriyorsa cevaplanması gereken temel soru, her bir sözcük için yapılan değerlendirmelerin nasıl birleştirileceğidir. İkili bağımsızlık elde etme modelinde iki sözcüğün birlikte görülme olasılığı her bir sözcüğün görülme olasılığının ortak olasılığı (joint probability) üzerinden hesaplanır. Dolayısıyla, değerlendirmede bir sözcüğün eklenmesi diğerini zenginleştirir. Eklenmiş sözcüğün varlığı, yok olduğu duruma göre daha gelişmiş bir durumdur. Modele adını veren ikili sözcüğü, bir dokümanın bir sorgu sözcüğünü içermesi ya da o sorgu sözcüğü ile ilgili olması durumunun, dokümanın sorgu sözcüğünü içermemesi ile birlikte ikili bir sonuç uzayı oluşturmasıdır. Bu yaklaşım sözcük sayısının ikiden fazla olduğu durumlarda kısıtlayıcı bir nitelik taşımaktadır.

Sözcüklerin birbirinden bağımsız olduğu kabul edilip, her sözcüğün bir doküman içerisinde bulunma olasılığı bağımsız olay olarak değerlendirilirse, bağımsız olasılıklar kullanılarak tüm sorgu sözcükleri için sonuç olasılığı hesaplanır. Birden fazla terim için terim olasılıklarının bağımsız kabul edilip çarpıldığı bu yaklaşım maksimum entropi prensibinin uygulanma yöntemlerinden biridir. Maksimum entropi prensibi üç temel varsayım etrafında geliştirilmiştir (Kantor and Lee, 1986):

1. Bir veri seti içerisindeki dokümanların, her bir sorgu sözcüğü ile ilgili olma durumuna göre yapılan değerlendirmelerden meydana gelen çapraz çizelgedeki (contingency table) doküman dağılımı, veri setine özgüdür. Bu dağılış herhangi bir özel varsayıma uymaz.

2. Başlangıçta veri seti içerisindeki bir konu ile ilgili doküman yüzdesi bilinmektedir. İki sözcük birlikte değerlendirilirken, eğer eşanlamlı iseler birleşme özelliği kullanılarak olasılıklar toplanır ya da ortak olasılık hesaplanması yolu ile sonuç olasılık değeri elde edilir.

3. Son olarak, verilen çapraz çizelgedeki konu ile ilgili olan ve olmayan dokümanların dağılışı 1. ve 2. maddelerdeki kısıtlamalara bağlı kalmak şartıyla mümkün olduğunca rastgeledir.

2.2.3 Mantıksal modeller

Mantıksal bilgi elde etme modelleri bilgi kavramlarını mantık bakış açısı ile değerlendirir. Sorgu-doküman ilişkilerini mantıksal çıkarımlar yolu ile ifade etmeye çalışır. Hedef dokümanlar içerisinden başlangıç noktaları seçilerek bu başlangıç noktalarından mantıksal çıkarımlar sonucunda, verilen sorguya ulaşılmasını sağlayanlar bulunmaya çalışılır. Şekil 2.3'de klasik IR, mantıksal bakış açısı ile birlikte bir örnek üzerinde verilmektedir (Fuhr, 2009):

Klasik IR	**Mantıksal Model**
$d=\{t_1, t_2, t_3\}$	$d = t_1 \wedge t_2 \wedge t_3$
$q=\{t_1, t_3\}$	$q = t_1 \wedge t_3$
Bilgi elde etme: $q \subset d$?	Bilgi elde etme: $d \rightarrow q$?

Şekil 2.3 Sorgu-doküman ilişkilendirmesine klasik IR ve mantık bakış açısı (Fuhr'dan 2009).

2.2.4 Dil modelleri

Bilgi elde etme alanındaki dil modelleri, diğer modellerin gözardı ettiği bir problem üzerine yoğunlaşır. Bu sorun anahatlarıyla; dokümanların sözcükler yolu ile gösteriminin (indeksleme), mantıksal ya da olasılık bakış açısı ile irdelenmesi biçiminde ifade edilebilir. Mantıksal ifade ile d doküman, t doküman içerisinde geçen sözcükler olmak üzere P(d→t) ya da P(t→d)'nin sorgulanması biçiminde gösterilebilir. Dokümanın yazıldığı dile özel doğal dil işleme (natural language processing) teknikleri devreye girdiği için, bu modeller dil modelleri olarak adlandırılmıştır. Doküman-sorgu ilişkilendirmesine bu bakış açısı ile yaklaşır. Örneğin Zhai ve Lafferty modelinde doküman-sorgu ilişkisini ifade eden olasılık sözcükler üzerinden aşağıdaki şekilde gösterilmiştir (Fuhr, 2009):

$$P(d \to q) \approx \sum_{m} P(d \to m) P(m \to q) \qquad (2.2)$$

2.2.5 Rastgelelikten ıraksama (Divergence from randomness)

Dil modeli yaklaşımı temel alınarak olasılık modelleri türetilmesini sağlayan Rastgelelikten Iraksama (Divergence from Randomness) modeli temel bir çerçeve sunar. Doküman içerisinde geçen sözcüklerin dağılımını olasılık dağılışları ile karşılaştırarak sonuçlara varmaya çalışır. Bu alandaki modeller iki sınıfta incelenir (Fuhr, 2009):

- **Binom ve Bose-Einstein modelleri**. Bir sözcüğün, eşit büyüklükteki dokümanlardan oluşan bir kümedeki dağılımına dayanır. Temel varsayım sözcüğün her bir dokümanda görülme olasılığının birbirine eşit olmasıdır.

  Binom modelinde temel olay bir sözcüğün bir dokümanda görülmesidir.

  Bose-Einstein modeli ise bir grup sözcük n tane dokümana serpiştirildiğinde belli bir k dokümanında sözcüğün görülme olasılığına dayanır.

- **Bernoulli ve Laplace modelleri**. Bir sözcüğün bulunduğu dokümanlar içerisinde birden fazla bulunma olasılığı üzerine yerleşir. Modeller bir doküman bir sözcüğü içeriyorsa bu sözcüğün yeniden bu dokümanda yer alma olasılığının hiç yokken bulunma olasılığından daha yüksek olduğu

varsayımına dayanır. Bu varsayım doğrultusunda yerel ve global normalizasyon işlemleri uygulanarak yeni sözcük ağırlıkları hesaplanır.

2.2.6 Etkileşimli modeller

Etkileşimli IR modelleri kullanıcı-makine etkileşimini modellemeye çalışır. Henüz gözlemci-gözlenen ilişkisinin modellenmesini sağlayacak kurallara uygun (formal) yöntemler mevcut olmadığı için matematiksel temeli yoktur.

2.2.7 Vektör uzayları ve genelleştirilmiş bir model gereksinimi

Soyut bir bakış açısı ile birer nesne olarak temsil edilen bilgi parçacıkları sayısal değerlerle eşleştirilmek üzere Salton tarafından geliştirilen Vektör Uzayı Modeli'nde (Salton, 1989) birer vektör olarak ele alınır; bilgi, vektör uzayları kullanılarak modellenir.

Vektör uzayı modelindeki temel varsayım, sorguların ve dokümanların birbirinden bağımsız olduğudur. Aradaki ilişkiler cebirsel işlemlerle ortaya konur. Bu model kapsamında özellik birleştirme (feature combination) mekanizmaları için matematiksel yöntem arayışına girilir. Diğer modellerde ise koşullu olasılık kavramının, özellikleri biraraya getirmede merkezi bir noktaya yerleştiği görülür. Örneğin, dil modellerinde doküman metninden hesaplanan özellikler doküman bağ yapısından hesaplanan özelliklerle olasılık varsayımları ve olanakları kullanılarak birleştirilir. Benzer bir birleştirmeyi vektör uzayı modelinde yapmak için arka plandaki uzay varsayımlarını dikkate alan, bu varsayımlarla ve ilgili geometrik yapılarla uyumlu cebirsel yöntemlerin önerilmesi gerekmektedir.

Vektör uzayı vektörlerden oluşan bir kümedir. Kurallara uygun (formal) bir şekilde tanımlanmak istenirse bir V vektör uzayındaki vektörler aşağıdaki aksiyomları sağlar:

V vektör uzayındaki her $\underline{x}$ ve $\underline{y}$ vektör çifti için bir $\underline{x}+\underline{y}$ toplam vektörü mevcuttur ve vektör toplama işleminin temel özellikleri aşağıda listelenmiştir;

1. Değişme özelliği, $\underline{x}+\underline{y}=\underline{y}+\underline{x}$,

2. Birleşme özelliği, $\underline{x}+(\underline{y}+\underline{z})=(\underline{x}+\underline{y})+\underline{z}$,

3. V'deki her $\underline{x}$ vektörü için orijin adı verilen, $\underline{x}+\Phi=\underline{x}$ eşitliğini sağlayan tek bir Φ vektörü vardır,

4. V'deki her $\underline{x}$ vektörü için $\underline{x}+(-\underline{x})=\Phi$ eşitliğini sağlayan tek bir $-\underline{x}$ vektörü vardır.

V vektör uzayındaki her α skaler değeri ve $\underline{x}$ vektörü için $\alpha\underline{x}$ çarpım vektörü mevcuttur ve çarpma işleminin temel özellikleri aşağıda verilmiştir;

1. Skaler ile çarpımın birleşme özelliği, $\alpha(\beta\underline{x})= (\alpha\beta)\underline{x}$,

2. Her $\underline{x}$ için $1\underline{x}=\underline{x}$,

3. Skaler ile çarpımın vektör toplaması üzerine dağılma özelliği,

   $\alpha(\underline{x}+\underline{y})=\alpha\underline{x}+\alpha\underline{y}$,

4. Vektör ile çarpımın skaler toplama üzerine dağılma özelliği

   $(\alpha+\beta)\underline{x}= \alpha\underline{x}+\beta\underline{x}$.

Vektör uzaylarında, vektörler arasındaki ilişkiyi ortaya koyan temel cebirsel yöntem (metrik) iki vektör arasındaki açıyı veren Cosine'dir. Cosine, vektörlerin skaler çarpımı ve vektör uzunluğu kavramlarından yola çıkılarak oluşturulmuş bir metriktir. Vektör uzunluklarından bağımsızdır.

Vektörlerin skaler çarpımı (inner product) aşağıdaki şekilde tanımlanır:

$$\left(\underline{x},\underline{y}\right)=\sum_{i=1}^{n}\overline{x}_i y_i, \qquad \overline{x}_i,\ x_i\text{'nin karmaşık eşleniği.} \qquad (2.3)$$

$$\overline{\mathrm{a}+\mathrm{bi}}=a-bi$$

Vektörler arasındaki skaler çarpım işleminden kaynaklanan ve geometrik olarak vektörün uzunluğu olarak yorumlanan norm kavramı, bir $\underline{x}$ vektörü için aşağıdaki eşitlikle ifade edilir ve her zaman bir reel sayıdır.

$$\|x\|=\sqrt{\left(\underline{x},\underline{x}\right)} \qquad (2.4)$$

Vektör uzunluklarından vektörler arasındaki uzaklık kavramına erişilir. Dolayısıyla $\underline{x}$ ve $\underline{y}$ vektörleri arasındaki uzaklık aşağıdaki şekilde tanımlanabilir:

$$d(\underline{x},\underline{y}) = \|\underline{x}-\underline{y}\| = \sqrt{(\underline{x}-\underline{y},\underline{x}-\underline{y})} = ((x_1-y_1)^2 + \ldots + (x_n-y_n)^2)^{1/2} \qquad (2.5)$$

Vektör uzunluğu Cauchy-Schwartz eşitsizliğini sağlar:

$$\forall \underline{x},\underline{y}, |(\underline{x},\underline{y})| \le \|\underline{x}\|\|y\|. \qquad (2.6)$$

Bu eşitsizlik kullanılarak bir başka eşitsizlik

$$-1 \le \frac{(\underline{x},\underline{y})}{\|\underline{x}\|\|\underline{y}\|} \le 1 \qquad (2.7)$$

yazılabilir ve $\underline{x}$ ve $\underline{y}$ vektörleri arasındaki açı φ iken aşağıdaki eşitlik verilebilir:

$$(\underline{x},\underline{y}) = \|\underline{x}\|\|\underline{y}\|\cos\varphi, 0 \le \varphi \le \pi, \qquad (2.8)$$

Bu eşitliğin sağında yer alan cosφ Cosine'yi verir. (Rijsbergen, 2004).

Bilgi parçacıkları vektör uzayları ile modellendiğinde, oluşturulan algoritmalar çok boyutlu dizi (array) veri yapısını kullanır. Nesnelerden, daha özel gösterimlere doğru giderken bu gösterimlerin karşılığı olan mevcut veri yapılarında, bilgi temsil edilir ve bu veri yapıları üzerinde geçerli olan işlemler kullanılarak, bilgi işlenir.

Günümüzde gerçekleştirilen modelleme çalışmalarında izlenen yol genellikle bilginin nasıl bir veri yapısı üzerinde temsil edilmesi sorusu üzerinde yoğunlaşır. Nitekim mevcut veri yapılarını farklı uzaylara eşleştirme, birden fazla veri yapısını birleştirme, kesikli veri yapılarını sürekli uzaylara ya da operatörlere yakınsama, veri yapısında indirgeme ya da genişletme yapma, üzerinde bir ölçüde olgunlaşılmış çalışma alanlarıdır. Bilgiyi matematiksel olarak modelleme çalışmaları bilgiden veri yapısına ve veri yapısından matematiksel yapıya yapılan dönüşümler olmak üzere birbirini tamamlayan iki çalışma alanını kapsar.

Bir başka modelleme bakış açısı, bir nesne ya da bilgi parçacığını sayısal bir veriye dönüştürmek yerine, bilgi parçacığı ya da nesneyi belli aralıkta değerler

alabilen bir değişken olarak değerlendirmektir. Bu bakış açısı programlama dillerinde veri tiplerini yaratan bakış açısıdır. Değişkenler birarada değerlendirilmeye başlandığında veri tiplerinden veri yapılarına geçiş gerçekleşir. Ayrıca bilgi parçacıklarını veri tipleri olarak değerlendiren anlayış Zadeh'in (2005) bilginin belirsizliğini modellemeye çalıştığı genelleştirilmiş belirsizlik teorisinin temel bakış açısı ile örtüşmektedir. Bu modelleme yaklaşımı olasılık teorisine dayanmaktadır.

Metrik uzaylar bilgi modelleme çalışmalarında genel çerçeveyi sunar. Uygulamalar vektör uzayı, olasılık ve mantıksal modellerde karşımıza çıkar. Uygulaması olan bu modellerin tümünü kapsayacak metrik uzaylardan daha genel bir model arayışı olduğu görülmektedir. Yeni bir model, yeni bir bakış açısını gerekli gören bu arayış aşağıdaki gereksinimleri karşılamaya yöneliktir (Rijsbergen, 2004):

- Varolan IR modelleri (vektör uzayı, olasılık, mantıksal vb.) tek bir çerçeve altında tartışılamıyor, temsil edilemiyor.

- Bilginin temel özelliklerinin kurallara uygun (formal) bir analizi mevcut değil.

- Kurallara uygun (formal) çıkarımların yapılabildiği semantik düzeyde çalışabilen temel bir model mevcut değil.

- Metin tabanlı bilgi yanında her çeşit ortamdaki bilginin analizinin yapılmasına her zamankinden fazla bir gereksinim vardır.

Rijsbergen, IR sistemleri için mantık, olasılık ve geometrinin aynı çatı altında temsil edileceği bir model arayışı içindedir. Sözkonusu modeli kuantum mekaniğini kullanarak oluşturmaya çalışmaktadır. Ortaya koyduğu görüşler aşağıdaki şekilde özetlenebilir:

1. IR modellerinde kullanılan skaler çarpım işleminde reel sayılar baz alınmaktadır. Oysaki daha genel bir model için karmaşık sayı skaler çarpımı daha uygun görünmektedir.

2. Boolean mantık, IR işlemleri için uygun değildir.

3. Bilginin belirsizlik (uncertainty) özelliğini karşılamak üzere olasılık teorisi ile ilişkilendirilmiş geometrik uzay ve işlemlerin tanımlanması gerekmektedir.

Rijsbergen'in düşünceleri henüz uygulamaya konmamıştır. Varolan modelleri birleştiren bir üst soyutlama seviyesini temsil eden kurallara uygun (formal) bir yapı tanımlamaya çalışmıştır. Bilginin doğasına uygun bir geometrik uzay/yapı arayışı içerisinde olması nedeniyle Rijsbergen bakış açısı bu çalışma ile benzerlik göstermektedir.

2.3 Manifold Kavramı

Bu tez çalışmasında; bilgi modellerinde genel çerçeveyi sunan metrik uzayların, Öklidyen olmayan geometri bakış açısı doğrultusunda genişletilmesi yolu ile uzunluk metriği yanında kullanılabilecek diğer metriklerle bilginin daha özel olarak temsil edilmesi hedeflenmektedir.

Öklidyen olmayan geometri manifold kavramını beraberinde getirir. Bir manifold, metrik uzayın daha genel bir halidir. Manifold sürekli, türevlenebilir ve analitik fonksiyonların geçerli olduğu bir ortamdır (Shiga and Sunada, 2005). Analitik fonksiyonlar, türevlenebilir fonksiyonların özel bir hali iken, türevlenebilir fonksiyonlar da sürekli fonksiyonların belli koşulları (ardarda türevi alınabilirlik) sağlamış halidir. Türevlenebilirlik, kapsama ilişkisi açısından süreklilik ile analitiklik arasında yer alır. Manifoldlar türevlenebilirlik kavramını tanımlarında bulundurmaları nedeniyle daha önceki uzay varsayımına özelleştirme, bu ek özelliğin sağlanması sonucunda yerel uzayların birleştirilerek global bir ortam kavramına ulaşılması nedeniyle ise bir genişletme kazandırır. Dolayısıyla, bilginin bir manifold üzerinde modellenmesi daha genel bir çerçeve sağlarken manifold üzerindeki yerel bölgeler sayesinde özelleşmiş bilgiler de karşılanabilir.

Manifoldlarda uzunluk metriğine ek olarak eğrilik metriği karşımıza çıkar.

Manifoldlar üzerinde uzunluk metriğinin dayandığı temel varsayım, uzunluğun konumdan (position) bağımsız olmasıdır. Konumun sabitlenmesi miktarın (büyüklüğün) sabitlenmesine indirgenebilir; dolayısıyla n boyutlu bir manifoldda bir nokta n tane değere (x_1, x_2, x_3,…, x_n) karşılık gelir. Bir doğru, bu değerlerin x'in fonksiyonu olarak ifade edilmesiyle tanımlanır. Doğrunun

uzunluğunu hesaplamak için söz konusu x değerleri bir birim cinsinden ifade edilmelidir. Birim olarak her bir bileşendeki değişim (artma veya azalma) oranı kullanılır. Bileşenlerin değişim oranının doğru üzerindeki toplam etkisi değerlendirilmek suretiyle uzunluk tanımlaması yapılmış olur (Riemann, 1873). Konum da uzunluktan bağımsız olduğundan; uzunluğun tanımlanmasında konum kullanılabilir. Nitekim Newton mekaniğinde de benzer bir durum söz konusudur. Zamanın mutlak ve konumdan bağımsız oldugu varsayılarak, konum hesaplamasında zaman kullanılmıştır.

Manifoldlar yapılarındaki yüzeylere (surface) indirgenebilir ve genel yaklaşım karmaşıklığın daha az olduğu yüzeyler üzerinde çalışmaktır.

Bir yüzey dâhilinde uzunluk metrik olarak kullanılabilir. Yüzey içerisindeki doğrular uzunlukları açısından karşılaştırılabilir. Bu karşılaştırmanın anlamlı olabilmesi için, söz konusu doğru uzunluklarının, dışsal etkilerden etkilenmiyor olması gerekir. Doğru uzunlukları yüzeye uygulanan bükme işleminden etkilenmez fakat esnetme, germe işlemlerinden etkilenir.

Bir yüzey içerisinde kullanılabilecek diğer bir metrik eğriliktir. Yüzey üzerinde tanımlanan bir nokta ve bir yöndeki eğrilik, yüzeyi içeren manifoldun o nokta ve o yöndeki eğriliğine eşittir. Eğrilik; vektörel bir niteliğe sahip olan yönü nedeniyle, konumdan bağımsız değildir.

Düz manifoldlarda herhangi bir noktanın herhangi bir yönünde eğrilik 0'dır. Eğriliği 0 olan manifoldlar sabit eğriliğe sahip olanların özel bir hali olarak düşünülebilir. Bir noktadaki bütün yönlerde sabit eğrilik değerleri elde edilen manifoldlar için, eğrilik geçerli bir ölçüttür. Eğriliğin verilen bir noktanın tüm yönlerinde sabit olmadığı manifoldlarda henüz geçerli bir ölçüt yoktur. Belki toplam eğriliğin 0 olduğu söylenebilir.

Bilginin modellenmesinde uzunluk metriğinin kullanımına ek olarak eğrilik metriği de dikkate alınırsa içinde bulunulan yapıya ait daha fazla özellik gözönüne alındığı için daha iyi bir temsil gerçekleşebilir. Eğriliğin kullanılabilirliği (feasibility) konusunda bir değerlendirme yapmak gerekirse bilgiden veri yapısına ve veri yapısından matematiksel yapıya dönüşüm zinciri kurulabildiğinden altyapı olarak mümkündür. Söz konusu zincirin birleştirici halkası Gauss-Bonnet teoremidir (Ueno et al., 2003). Gauss-Bonnet teoremi bir yüzeyin toplam Gauss eğriliğinin o yüzeyin Euler karakteristiğinin 2π katı olduğunu ifade eder ve Euler

karakteristiği çizge veri yapısında hesaplanabilen bir metriktir. Çizgeden geometrik bir yapıya dönüşüm ise bir sonraki bölümde incelenecek jeodezik kavramı üzerinden gerçekleştirilebilir.

2.4 Jeodezik Uzunluklar

Çizgelerden manifoldlara geçiş, eğriliği dikkate alan uzunluk (jeodezik) kavramı yardımı ile gerçekleşir. Çizgelerde düğümleri (vertex) birbirine bağlayan kenarlar (edge) jeodezikler ile eşlenir. Jeodezik iki nokta arasındaki en kısa eğridir. Bir başka deyişle ivmelenmeyen bir parçacığın izlediği yoldur (Rowland and Weisstein). Bir jeodeziğin Öklid kirişinden ayrılma ya da kıvrılma derecesi bölgesel eğrilik (sectional curvature) kavramı ile karşılanır. Dolayısıyla iki düğümü birbirine bağlayan jeodezik uzunluk (çizgenin kenar ağırlıkları) Öklidyen uzunluk ve bölgesel eğrilik değerleri üzerinden hesaplanır (Robles-Kelly and Hancock, 2007).

Manifold üzerindeki iki nokta arasındaki enerjinin ifade edilişinde, jeodezik demeti için tanımlanmış Jacobi vektör alanı ve bu alanın eğrilik tensörü ile ilişkisi kullanılır. Eğriliğin eğrilik tensörü üzerinden en genel formülü aşağıda verilmektedir (Wrede, 1972):

$$\kappa = \kappa\left(x; X_a, Y_a\right) = \frac{R_{ijkl} X_a^i Y_a^j X_a^k Y_a^l}{G_{pqrs} X_a^p Y_a^q X_a^r Y_a^s} \quad \left(G_{pqrs} \equiv g_{pr} g_{qs} - g_{ps} g_{qr}\right) \tag{2.9}$$

Formülde κ eğriliği, X_a ve Y_a bir a noktasındaki doğrusal bağımsız vektörleri, R_{ijkl} eğrilik tensörünü ve G_{pqrs} ise metrik tensörünü ifade etmektedir. Riemann eğriliği sadece konuma değil her noktada seçilen yön çiftine de bağlıdır.

Eğrilik tensörü iki boyutta değerlendirildiğinde (n=2) yönden bağımsız duruma gelip bölgesel eğrilik kavramı ile karşılanır.

$$\kappa = \frac{R_{1212}}{g_{11} g_{22} - g_{12}^2} \equiv \frac{R_{1212}}{g} \tag{2.10}$$

Riemann eğriliği 2-boyutta g_{ij} ve türevlerine bağlı olup seçilen doğrusal bağımsız vektör çiftinden bağımsızdır.

Çizge-manifold eşleştirme yöntemlerinde manifold, sabit bölgesel eğriliğe sahip olarak kabul edilip eğrilik tensörünün ifade edilen bu özelliğinden yararlanılır.

X_a ve Y_a doğrusal bağımsız vektörler olmak üzere bu iki vektörün temsil ettiği yüzeyin eğriliği aşağıdaki gibidir:

$$\kappa_\delta = \kappa(X_a, Y_a) = \frac{(R(X_a, Y_a)Y_a \mid X_a)}{\|X_a\|^2 \|Y_a\|^2 - (X_a \mid Y_a)^2} \qquad (2.11)$$

Bir γ parametrik eğrisi bir M manifoldu üzerinde

$$\gamma : t \in [\alpha, \beta] \mapsto M \qquad (2.12)$$

ifadesi ile tanımlanmak üzere ivme vektör alanı 0 ise ya da hız vektör alanı eğri boyunca paralel ise jeodezik özellikleri taşır (Mukerjee):

$\nabla_{\gamma'}^{\gamma'} = 0$ ∇: kovaryant türev, γ': hız vektör alanı. (2.13)

Jeodezik, bir Y Jacobi vektör alanı içerisinde ise aşağıdaki Jacobi eşitliğini sağlamalıdır:

$$\nabla_t^2 + R(\gamma', Y)\gamma' = 0 \qquad (2.14)$$

γ jeodeziği ve Y Jacobi vektör alanının temsil ettiği yüzeyin eğriliği aşağıdaki şekilde ifade edilir:

$$\kappa_\delta = \kappa(Y, \gamma') = \frac{(R(Y, \gamma')\gamma' \mid Y)}{\|Y\|^2 \|\gamma'\|^2} \qquad (2.15)$$

Bir manifold üzerindeki p_u ve p_v noktalarını birleştiren jeodezik üzerindeki enerji, yüzey eğriliği ve Jacobi eşitlikleri kullanılarak aşağıdaki eşitlik ile ifade edilir (Robles-Kelly and Hancock, 2007):

$$\begin{aligned} \varepsilon(p_u, p_v) &= \int_\gamma \left|\gamma' + \nabla_t^2 Y\right|^2 dt \\ &= \int_\gamma \left|\gamma' - \kappa(\gamma', Y)Y\right|^2 dt \end{aligned} \qquad (2.16)$$

Bu eşitlikten manifold üzerindeki iki nokta arasındaki enerjinin (jeodezik uzunluğun) Öklidyen uzunluk ve sabit bölgesel eğrilik değerlerine bağlı olduğu görülebilir.

Manifold boyunca görülen sabit eğrilik değerleri güçlü global özelliklerin göstergesidir. Ayrıca sabit eğrilik değerleri, içinde bulunulan manifoldun sınırlarının ve büyüklüğünün tahminlenmesini sağlar. Pozitif κ sabit eğrilik değerleri görülen bir manifold sonludur ve çapı $\pi/\sqrt{\kappa}$ 'ya eşittir. κ negatif ise manifold sonsuzdur (Jonckheere and Poonsuk, 2004).

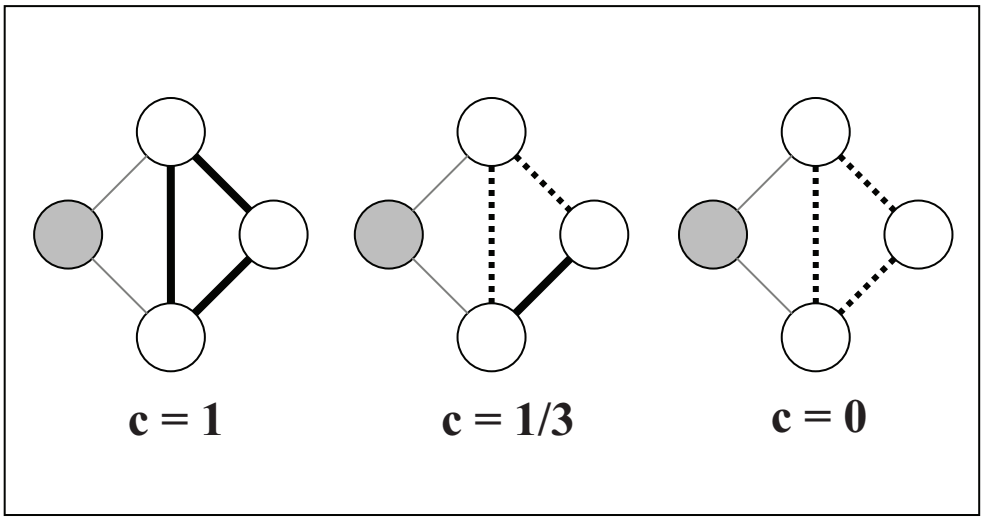
Bölgesel eğrilik değerleri hesaplamasında kümelenme katsayısı ve Alexandrov açıları kullanılır ve yöntemler birbiri ile tutarlı sonuçlar üretir. Bölgesel eğrilik değerlerinin dağılışı incelenerek global sonuçlar çıkarmak mümkündür.

Kümelenme katsayısı bir çizgede bir düğümün komşuluğundaki düğümler arasındaki toplam kenar sayısının mümkün olabilecek en fazla kenar sayısına bölümüdür ve aşağıdaki şekilde formüle edilebilir (Watts and Strogatz, 1998):

$$C_i = \frac{2\left|\{e_{jk}\}\right|}{k_i\left(k_i - 1\right)} : v_j, v_k \in N_i, e_{jk} \in E. \qquad (2.17)$$

e, iki düğüm (vertex) arasındaki kenar sayısını verirken k çizgedeki toplam düğüm sayısıdır.

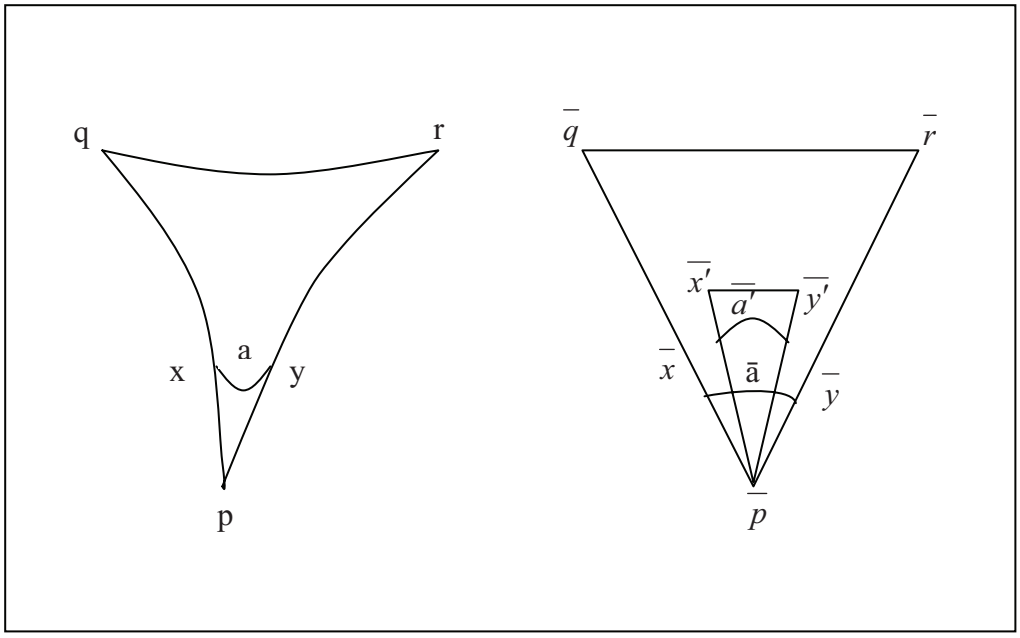
Şekil 2.4'te ise bir düğümün farklı komşuluk yapılarına göre elde edilen kümelenme katsayısı değerleri verilmektedir (Wikipedia).



Şekil 2.4 Yönsüz bir çizgede kümelenme katsayısı örnekleri (Wikipedia'dan).

Kümelenme katsayısı skaler eğrilik değerlerini tahminler. Kümelenme katsayısının 0 değeri negatif eğriliği, 1 değeri pozitif eğriliği, 0,5 değeri ise 0 eğrilik değerini karşılar (Lou, 2009).

Alexandrov açıları jeodezik üçgenlerin açılarıdır. Şekil 2.5'te solda bir jeodezik üçgen ve üçgenin a Alexandrov açısı görülmektedir. Sağda bu üçgene ait karşılaştırma üçgeni verilmiştir.



Şekil 2.5 Jeodezik üçgen ve karşılaştırma üçgeni.

Kosinüs teoremi (law of cosines) üçgenlerde bir açının kosinüsünün üçgenin kenarları cinsinden karşılığını verir. Kosinüs teoremi, kenarları p,q,r ve p kenarını gören açısı α olan bir üçgen için aşağıdaki şekilde ifade edilebilir:

$$p^2 = q^2 + r^2 - 2qr\cos(\alpha)$$
$$\cos(\alpha) = \frac{q^2 + r^2 - p^2}{2qr} \qquad (2.18)$$

Verilen jeodezik üçgen ve karşılaştırma üçgeni (Bkz. Şekil 2.5) aşağıdaki açı ve kenar uzunluğu sıralamalarını doğrular:

$$a = \overline{a'}$$
$$\overline{a'} \le \bar{a}$$
$$d(x,y) = d(\overline{x'}, \overline{y'}) \quad \text{(karşılaştırma üçgenleri)}$$
$$d(x,y) \le d(\bar{x}, \bar{y}) \quad \text{(CAT}(\kappa)\text{-eşitsizliği)}$$

Kosinüs teoremi, jeodezik üçgenler ve bu üçgenlere ait karşılaştırma üçgenleri (comparison triangles) ile birlikte değerlendirildiğinde uzunluk sıralamasının açı sıralaması ile örtüştüğü görülür:

$$d(\overline{x'}, \overline{y'}) \le d(\bar{x}, \bar{y}) \Leftrightarrow \overline{a'} \le \bar{a} \quad \text{(Kosinüs teoremi)} \qquad (2.19)$$

Bu nedenle Jeodezik üçgenlerin Alexandrov açıları toplamının 2π değerinden farkı $2\pi - \sum_k \bar{\alpha}_k$ iyi bir eğrilik tahminlemesi verir.

Çizgenin bir düğümü etrafındaki eğrilik değeri, o düğümün köşesi olduğu üçgenlerde Alexandrov açıları toplamının 2π'den farkının üçgenlerin alanlarına bölümü sonucu elde edilen sayının toplam üçgen sayısına bölünmesi ile hesaplanır. Bu hesaplamaya ilişkin eşitlik aşağıda verilmektedir:

$$\kappa(a) = \frac{1}{\binom{\deg(a)}{3}} \sum_{\sigma \supset a} \frac{2\pi - \sum_{i=1}^{3} \overline{\alpha_k}}{\sum_{i=1}^{3} A(\overline{pqr})} \qquad (2.20)$$

Verilen eşitlikte σ çizgedeki kenarlardan oluşan bölgeyi, a jeodezik üçgenin köşesi olan düğümü $\bar{p}$, $\bar{q}$ ve $\bar{r}$ değerleri ise düğümler arasındaki uzunlukları temsil etmektedir.

Eğrilik değerleri ile Öklidyen uzunluklara bağlı olan jeodezik uzunluklar, içinde bulunulan manifoldun yapısına bağlı olarak farklı yöntemlerle hesaplanır. O yüzden jeodezik hesaplamalarda bir geometrik yapı varsayımı verilmelidir. Eğrilik değerlerinin belirlediği geometri, jeodezik hesaplama yöntemini de belirler.

Manifoldlar eğrilik değerlerine göre sınıflandırılır. Negatif eğrilik değerlerine sahip olanlar hiperbolik, pozitif eğrilik değerleri görülenler parabolik ve 0 sabit eğriliğin görüldüğü manifoldlar düzdür. Her bir sınıf için farklı bir jeodezik uzunluk hesaplama yöntemi geçerlidir.

2.5 Geleneksel Modellerde Öklidyen Olmayan Yaklaşımlar

Hui-Fang et al. (2008) jeodezik uzunlukları metin tabanlı analiz alanında kullanan ilk araştırmacılardır. Çalışmalarında jeodezik uzunluğu sorgu tabanlı cümle elde etme bağlamında kullanıp Cosine ile karşılaştırmaktadırlar. Önerilen yöntemde metinden ve sorgudan çıkarılan cümlelerden bir çizge oluşturulur. Bu çizge üzerinde yerel komşuluğu tanımlayan bir eşik değeri tayin edilir. Seçilen iki cümle arasındaki uzaklık belirlenen eşik değerinden küçükse bu iki cümleyi temsil eden düğümler arasında doğrudan bağlantı kurulur. Jeodezik uzunluklar cümlelerin oluşturduğu çizge üzerinde kaynak ve hedef düğümler arasındaki mesafenin en kısa yol (shortest path) algoritmaları ile bulunması sonucunda hesaplanır. Bu şekilde hesaplanan jeodezik uzunluklar ile oluşan derecelendirmeler ve doğruluk oranlarına ait grafikler Cosine durumunda elde edilenler ile karşılaştırılır. Sonuçlar eşik değerini ifade eden parametrenin belli değerleri için jeodezik uzunlukların Cosine'den daha iyi olduğunu gösterirken diğer değer aralıklarında Cosine performansının yinelendiği gözlenir.

Goth et al (2005) tarafından ortaya konan hiperbolik IR modelinde Öklidyen olmayan bir yaklaşım mevcuttur. Jeodezik uzunlukları hesaplayabilmek için uzaydaki noktalar belli bir geometrik yapı üzerinde varsayılır. Seçilen geometrik yapı hiperbolik küredir. Sorgu vektörü hiperbolik kürenin merkezi olarak kabul edilirken diğer dokümanlar kürenin merkezine olan uzaklıkları dikkate alınarak derecelendirilir. Yaklaşımdaki temel nokta, Öklidyen olmayan yöntemlerin belirleyici özelliğinin konum özelleştirmesi olduğudur. Zira eğrilik konuma bağımlıdır. Çalışmanın önemli bir sonucu; hiperbolik kürenin yarıçapının parametre olarak alınması durumunda, yarıçapa bağlı olarak değişen hiperbolik uzunluk değerlerine dayalı derecelendirmelerin, klasik benzerlik ölçülerinin

ağırlıklandırma mekanizmaları ile birlikte oluşturdukları sıralamalar ile örtüşmesidir. Özetle; ağırlıklandırma mekanizmalarının işlevi de hiperbolik benzerlik ölçüsü tarafından doğal olarak karşılanmaktadır.

Bir başka çalışmada (Xiao and Hancock, 2005) jeodezik, Öklidyen uzunluk ve bölgesel eğrilik değerlerine bağlı bir maliyet ile ilişkilendirilir. Bölgesel eğrilikler bir jeodeziğin Öklid kirişinden ayrılma derecesi ile belirlenir. Dolayısıyla Öklidyen ve jeodezik uzunlukların ölçüldüğü bir ortamda bölgesel eğrilik değerleri hesaplanabilir. Aynı şekilde bölgesel eğrilikler ve Öklidyen uzunluklar biliniyorsa, jeodezik uzunlukları hesaplamak mümkündür. Lou (2009), kümelenme katsayısı değerlerinin bölgesel eğrilik değerlerinin tahminlenmesinde kullanılabileceğini ifade eder.

Jeodezik uzunluklar metin bağlamında tek bir özellik (feature) uzayı dâhilinde kullanılabileceği gibi, birden fazla özellik uzayını birleştirici bir yöntem olarak da önerilebilir. Bu tez çalışmasında jeodezik uzunlukların, özellik uzaylarını birleştirici fonksiyonu test edilmektedir. Metin bağlamından elde edilen benzerlik ölçüsü dokümanlar arasındaki bağ bilgisi ile birlikte değerlendirilmeye çalışılır. Dolayısıyla metin bağlamında, bağ tabanlı temel yaklaşımların açıklanması faydalı olacaktır.

PageRank (Page et al., 1999) algoritması dokümanların derecelendirilmesinde doküman kümesine ait global bağ çizgesini kullanır. Bu çizgeyi baz alarak internet üzerindeki sörf davranışını modellemeye çalışır. Kleinberg (1999) tarafından geliştirilen HITS algoritması önemli dokümanları ifade etmek üzere otorite (authority) kavramını ortaya koyar. Buna göre bir dokümandan bir başka dokümana bağ verilmesi, o dokümana otorite verilmesi ile eşdeğerdir. Aynı zamanda kimden otorite alındığı da son derece önemlidir. Otorite dokümanlarının varlığına ek olarak otorite vermede yetkin merkezler (hub) bulunmaktadır. Algoritma bu merkez-otorite örüntüsü etrafında geliştirilmiştir ve metin tabanlı yöntemlerle yapılan filtrelemeler sonucunda elde edilen doküman kümesi üzerinde uygulanır ve global bağ çizgesi yerine yerel bağ çizgesi kullanılır.

Bağ bulgularını metin tabanlı skorlarla birleştirmek üzere dil modelleri de yoğun olarak kullanılmaktadır. Kamps ve Koolen'in (2009) deneysel çalışmaları, yerel bağ öncellerinin (prior) global olanlara oranla daha yararlı olduğunu ortaya koymaktadır. Hatta yerel ve global bağ bilgileri arasında ağırlıklandırma yapan mekanizmalar daha iyi sonuçlar vermektedir. Bu noktadan hareketle, global ve

yerel bağ bilgilerini birlikte değerlendiren, bu bilgiler arasında uzlaşma sağlayan yaklaşımların iyi sonuçlar vereceği öngörülmektedir.

Bağ bulgularını içerik tabanlı sınıflandırıcılarla birleştirdikleri çalışmalarında Calado et al (2003) özellik birleştirici yaklaşımların başarısının birleştirilecek her bir bulgu kaynağına atanacak önem derecesine bağlı olduğunu ifade etmektedir. Daha fazla güvenilir olan bilgiye daha fazla ağırlık verilmelidir. Bu alandaki temel problemin, birleştirilecek bilgilerin her birine uygun ağırlıkları verecek bir yöntemin bulunması olduğu hususunda karar kılınmıştır. Hedef, bağ tabanlı ve içerik tabanlı bilgileri biraraya getirecek alternatif yollar ile birlikte her birine uygun ağırlıkları tayin edecek otomatik yöntemleri bulmak olmalıdır.

Yang (2001) ise, bağ analizi ve bağ bilgilerini diğer tip bilgilerle birleştiren yöntemler (füzyon) konusundaki araştırma sonuçlarının bir fikir birliği ortaya koymadığına dikkat çekmektedir. Bazı çalışmalar birden fazla özelliğin birleştirildiği yaklaşımların sonuçları iyileştirdiğini ifade ederken diğerlerinin sonuçları füzyonun bilgi elde etme performansını düşürdüğünü ortaya koymaktadır. Yang'a göre temel soru füzyonun başarısız olduğu durumları izah edebilmektir. Test derlemlerine özgü karakteristikler, bağ analizindeki başarısızlıklar, füzyon formülündeki yetersizlikler birlikte ya da ayrı ayrı etkili faktörler olabilir. Yang, gelecekteki füzyon çabalarının farklı bilgi elde etme yöntemlerinin füzyon potansiyelini gerçeğe dönüştürecek füzyon formülünün keşfedilmesi üzerine odaklanacağına inanmaktadır.

Bir sonraki bölümde, vektör uzayı modeli üzerine yerleşen Öklidyen olmayan geometri kaynaklı jeodezik uzunluk benzerlik ölçüsü ele alınmaktadır. Tezdeki uygulama kapsamında kullanılan jeodezik uzunluk hesaplama yöntemi de açıklanmaktadır.

3. ÖNERİLEN ÖKLİDYEN OLMAYAN YAKLAŞIM

Bilginin semantik analizi ve modellenmesi ilk bölümde de ifade edildiği gibi büyük bir hedeftir. Farklı soyutlama seviyelerindeki bilgi özelliklerinin garanti edilmesini gerektirir. Bu hedefi destekleyici yeni bir semantik analiz yöntemi yaratmak üzere varolan bilgi modelleme çalışmaları Öklidyen olmayan bir bakış açısı ile değerlendirilebilir. Bu değerlendirme bağlamında, öncelikle üzerinde çalışılan bilgiye (veri setine) ilişkin geçerli olan varsayımların ortaya konması gerekir. Bu varsayımlar doğrultusunda aşağıdaki adımlardan oluşan geometrik bir metodoloji takip edilir (Dubrovin et al., 1991):

- Veri setinin içinde bulunduğu geometrik uzayın belirlenmesi.

- Veri setinin temsil ettiği geometrik yapının tanımlanması, cebirsel gösteriminin yapılması ve temel özelliklerin ortaya konması.

- Geometrik yapıyı temsil eden içsel (intrinsic) metriklerin belirlenmesi ve hesaplanması.

Verilen geometrik metodolojinin ortaya koyduğu yaklaşım, gerçek metriklerin içsel (intrinsic) ve dışsal (extrinsic) olarak sınıflandırılmasıdır. İçsel metrikler tanımlanan geometrik yapı içerisindeki değişmezler iken dışsal olanlar geometrik yapı içerisinde farklılık yaratmadığı halde, bu yapı dışındaki uzaydan bakıldığında ayırt edilebilenlerdir. Bir eğri bu bağlamda değerlendirilecek olursa, eğri için uzunluk içsel bir metrik iken eğrilik dışsaldır çünkü eğri üzerinde gidilirken uzunluk değişiminden bahsedilir, eğrilik algılanmaz. Eğri üzerindeki iki nokta ancak uzunluk metriği ile karşılaştırılabilir. Bu kavramsal açıklamayı yüzeyler için yorumlamak istersek, bir yüzey boyunca ilerlerken eğriliğin değiştiğini farkedebiliriz ve uzunluğu eğriliği dikkate alarak değerlendirmek zorunda kalırız. Yüzey üzerindeki iki noktayı eğrilikleri açısından karşılaştırabiliriz. Benzer şekilde, hacmi olan bir büyüklüğün akışkanlığından bahsedilebilir. Kısacası metriklerin kendi içerisinde bir sınıflandırması vardır ve içinde bulunulan geometrik yapıya göre uygunluklarının ele alınması gerekir. Örneğin iki doküman vektörünün uzunlukları kolayca metrik olarak kullanılabilirken eğriliği değerlendirmeye almak istersek bir yüzey kapsamında ölçmek daha makul olacaktır. Çünkü birden fazla dokümanın yerel bağlantısallık derecesini gözönüne almak ve normalize etmek daha güvenilir sonuçlar alınmasında fayda sağlayacaktır.

Yukarıda tanımlanan geometrik metodoloji gözönüne alınarak, üzerinde çalışılacak bir veri seti seçilir. Seçilen veri seti için uzay ve geometrik yapı varsayımı yapılır. Son olarak, kullanılacak metriğin hedefi belirlenir ve bu hedef doğrultusunda metrik değeri hesaplanır.

Gerçekleştirilen çalışmada metin tabanlı, bağlı bir veri seti seçilmiştir. Veri seti için yapılan uzay ve geometrik yapı varsayımı küresel geometridir. Küresel yapı üzerinde hesaplanan jeodezik uzunluklar metrik olarak kullanılmıştır ve hedef, bu metriği **dokümanların anlamsal yakınlıklarını ölçmek** üzere kullanmaktır. Dolayısıyla önerilen metrik aynı zamanda bir semantik analiz yöntemidir.

Önerilen semantik analiz yöntemi ya da Öklidyen olmayan benzerlik ölçüsü, vektör uzayı modeli kapsamında kullanılan metriklerin üzerine yerleşmekte, başka bir ifade ile mevcut altyapıyı kullanmaktadır.

Dokümanlar arasındaki benzerlik geleneksel bakış açısı ile ölçülmek istendiğinde, veri setinden dokümanları tanımlayıcı özellikler (feature) çıkartılmaya çalışılır. Bu özellikler doğal olarak farklı soyutlama seviyelerinde yer alır. Düşük seviyedeki özellikler doğrudan hesaplanabilir olanlardır; doküman uzunlukları, doküman içerisindeki sözcük frekansları buna örnek olarak verilebilir. Düşük seviyedeki özelliklerin varlığı, yüksek seviyedeki özelliklerin hesaplanmasında kullanılacak olmaları nedeniyle kritiktir.

Dokümanlar arasındaki anlamsal yakınlık, yüksek seviyede bir özelliktir ve düşük seviyedeki özellikler üzerinde karmaşık (semantik) kurallar uygulanarak ölçülür. Semantik analiz yöntemlerinde uygulanan karmaşık kurallar genellikle cebirsel kurallardır. Cebirsel dönüşümler soyutlama seviyeleri arasında geçiş yapılmasını sağlayan ideal yöntemlerdir. Benzer şekilde doküman vektörleri arasında benzerlik ölçümü vektörler arasındaki uzaklık üzerinden yapılır. Olağan uzaklık metriği, konumdan bağımsızdır çünkü iki nokta (vektör) arasındaki uzaklık her iki noktanın bir koordinatındaki aynı miktarlı bir ötelemeden etkilenmez. Bir başka deyişle, klasik Öklid uzaklığı, doküman uzayına ait topolojiyi dikkate almaz, uzay; eğriliğin 0 olduğu düz bir uzay olarak kabul edilir. Öklidyen olmayan geometrinin sunduğu eğrilik metriği bu bağlamda veri noktaları hakkında ek bir bilgi sağlayabilir zira eğrilik uzaydaki konuma bağlıdır.

Öklidyen olmayan benzerlik ölçüsünün önerilmesini sağlayan bakış açısı dokümanların anlamını daha iyi temsil edebilecek geometrik bir yapı arayışıdır. Bir veri setine ait dokümanların anlamını daha iyi bir şekilde temsil etmek üzere doküman uzayının geometrik yapısını yakalayan bir genelleştirmeye gereksinim vardır. Bu genelleştirme geometrik olarak uzaklık, açı, hacim ve eğrilik kavramlarını mümkün olduğunca içermelidir. Listelenen bu kavramlar koordinat sistemlerinden bağımsız gösterimlerdir ve bu kavramları içeren genelleştirilmiş metrik geometride g (Riemann) metriği olarak adlandırılır (Dubrovin et al., 1991). g metriği, içinde bulunulan uzayın kendine has özelliklerine göre farklı şekilde formüle edilir. g metriği belli bir uzay varsayımı doğrultusunda iki nokta arasındaki uzaklığın hesaplanma yöntemini ortaya koyan bir fonksiyondur. g metriğine bağlı uzunluk gösterimi aşağıdaki formülde verilmektedir:

$$ds^2 = g_{ij} dx^i dx^j \qquad (3.1)$$

g metriği daha soyut bir şekilde, bir manifold içerisindeki iki nokta arasındaki uzaklık ölçülmek istendiğinde o noktalardaki tanjant vektörlerinin kendi aralarındaki skaler çarpımlarını ifade eden matris olarak tanımlanabilir. Bu tanıma göre ξ ve η bir noktadaki tanjant vektörlerini ifade ederse g metriği aşağıdaki matris ile temsil edilebilir:

$$g_{ij} = \begin{bmatrix} \langle \xi, \xi \rangle & \langle \xi, \eta \rangle \\ \langle \eta, \xi \rangle & \langle \eta, \eta \rangle \end{bmatrix} \qquad (3.2)$$

Koordinat eksenlerinin birbirine dik olduğu Öklid geometrisinde tanjant vektörleri birbirine dik olduğu için skaler çarpımları 0'dır. Öklid geometrisinde g metriği Kronecker delta ile karşılanır ve aşağıdaki şekilde gösterilir:

$$\delta_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad (3.3)$$

Bu çalışmada önerilen Öklidyen olmayan metrik de belli bir uzay varsayımına göre oluşturulan bir uzunluk hesaplama yöntemidir ve g metriğinden esinlenerek oluşturulmuştur. Eğrilik değerleri ile zenginleştirilmiş uzunluk hesaplamaları (jeodezik uzunluklar) doküman uzayının kendine has özelliklerini yansıtmada fayda sağlayacaktır.

Teoride, içinde bulunulan uzaya ve geometrik yapıya özgü uzunluk hesaplama yöntemleri geliştirilmiştir. Pratikte bu yöntemlerin kullanımında iki temel güçlükle karşılaşılır:

1. Veri setlerinin gerçekte temsil ettiği yapıyı ve bu yapıyı içeren uzayı tahminlemek oldukça zordur.

2. (1).maddede belirtilen tahminlemeyi yapmak yerine veri seti belirli bir uzay ve yapı varsayımı doğrultusunda analiz edilebilir. Ancak; eşleştirme iyi bir eşleştirme olmayabileceğinden, yanılma payı yüksektir.

3. Yapılan eşleştirmeler doğru olsa dahi teorik hesaplamanın parçası olan her parametreyi ve katsayıyı gerçekte ölçmek mümkün değildir.

Mevcut vektör uzayı modelini bu bakış açısı ile değerlendirmeye çalışalım.

Vektör uzayı modelinde temel uzay varsayımı Öklidyendir. Dokümanlar, içerisindeki her sözcük bir koordinata eşlenecek şekilde çok boyutlu vektörler olarak temsil edilir. Koordinat eksenlerinin birbirine dik olduğu varsayılmaktadır. Başka bir deyişle her sözcüğün birbirinden bağımsız olduğu kabul edilmektedir. Oysa sözcükler arasında anlamsal ilişkiler vardır. Kullanılacak model bu anlamsal ilişkileri ortaya koymalıdır. Bu noktadan hareketle modelin bu eksikliğini giderecek tamamlayıcı yaklaşımlar ortaya çıkmıştır. Bunlardan, gizli semantik indeksleme (latent semantic indexing) yaklaşımında doküman vektörlerinin oluşturduğu terim-doküman matrisinin, yapıtaşı olan daha küçük bir matrise indirgenmesi hedeflenmektedir (Deerwester et al., 1990). İndirgeme işlemleri tekil değer ayrıştırma (singular value decomposition) yöntemi ile gerçekleştirilir. Temel fikir, elde edilen daha küçük ranka sahip matrisin sözcükler arasında varolan anlamsal ilişkileri daha iyi ifade edeceği, bir başka deyişle elde edilen matriste birbirine yakın anlamlı sözcüklerin aynı koordinata yakınsayacağıdır. Böylece yeni matristeki eksenler diklik varsayımına daha uygundur.

Vektör uzayı modelinde gözönüne alınması gereken önemli bir husus, doküman vektörleri oluşturulurken kullanılan ağırlıklandırma ölçüleridir. Zira ağırlıklandırma ölçüleri her boyut için aynı (uniform) olmayan bir çarpıklık (distortion) yaratır ve klasik dik bir uzay yerine eğriliğin gözlendiği Öklidyen olmayan bir uzay ortaya çıkar.

tf-idf en temel ağırlıklandırma ölçüsüdür. tf (term frequency) doküman içerisinde geçen terim frekansıdır. idf (inverse document frequency) toplam doküman sayısının terimi içeren doküman sayısına oranını ölçen bir değerdir ve temel prensip olarak hemen hemen her dokümanda geçen terimlerin bir dokümanın anlamına katkısının az olacağı varsayımıyla tf değerini normalize etmek için kullanılır. Aşağıda, temel idf hesaplaması verilmektedir (Jones, 1988):

$idf = \ln[N_{Docs} / (docFreq + 1)] + 1,$
N_{Docs} : Toplam doküman sayısı, $docFreq$: Terimin geçtiği doküman sayısı.

Genel olarak Öklidyen geometri Öklidyen olmayan geometrinin özel bir halidir. Bu yüzden doküman uzayını Öklidyen bir bakış açısı yerine Öklidyen olmayan bir bakış açısı ile modellemek, varolan tüm varsayımları sorgulayıp bir üst soyutlama seviyesine çıkmayı gerektirdiği için oldukça güçtür. Dolayısıyla bu çalışma kapsamında Öklidyen olmayan bir yaklaşımı uygulamak üzere alternatif olarak vektör uzayı modeli üzerinde Öklidyen olmayan benzerlik ölçüleri düşünülmüştür. Bu bakış açısı modelde ağırlıklandırma mekanizmaları ile benzerlik ölçülerinin birlikte kullanımının Öklidyen olmayan bir yapı oluşturacağı düşüncesi ile de bütünlük taşımaktadır. Bu çalışma ile Öklidyen olmayan bir benzerlik ölçüsü geliştirilmekte ve bu yeni bakış açısının getireceği katkı araştırılmaktadır.

Mevcut modellerde doküman uzayının Öklidyen olması bir varsayımdır, matematiksel bir tahminin sonucu değildir. Dolayısıyla öyle olduğunu gösteren bir delil varolmadığına göre, uzayı bir manifold olarak düşünmek daha doğal ve makuldur. Benzer düşünceyle (He et al., 2004) doküman uzayını doğrusal olmayan (non-linear) yöntemler kullanarak altuzaylara indirgemek geçerli bir yaklaşım olmuştur. Bununla birlikte doğrusal bir gösterimden doğrusal olmayan daha küçük gösterimlere geçmenin ötesinde daha önce de belirtildiği gibi ilk gösterimi bir manifold olarak görmek bakış açısında ve uygulanacak yöntemlerin değerlendirmesinde bir genişletme sağlayabilir. Çünkü bütün diğer yöntem ve hesaplamalar varsayımlar üzerine yerleşir ve varsayımlar dikkate alınarak şekillendirilir. Ayrıca bir yapıdan indirgeme yapmak bir ölçüde bilgi kaybına uğramakla eşdeğerdir.

Vektör uzayı modeline bir alternatif olarak önerilen hiperbolik IR modelinde doküman uzayının hiperbolik bir uzay olduğu kabul edilir (Goth and Skrop,

2005). Vektör uzayı modelinin aksine doküman vektörleri birbirine dik değildir, hiperbolik uzay varsayımı doğrultusunda yerleşir, konumlar önem kazanır. Hiperbolik IR modelinde konumlar özelleşir çünkü dokümanlar hiperbolik küre geometrik yapısı üzerine yerleştirilmeye çalışılır. Sorguyu ifade eden konum (q vektörü) hiperbolik kürenin merkezine çekilir, diğer dokümanlar bu kürenin merkezine olan uzaklıkları (hiperbolik uzaklık) kullanılarak değerlendirilir. Sonuç olarak, Öklidyen olmayan bir yaklaşım uygulanmak isteniyorsa metrik uzayda birer eleman olarak temsil edilen noktaların varsayılan öklidyen olmayan yapı yolu ile daha da özelleştirilmesi, farklılaştırılması gerekmektedir.

3.1 Jeodezik Benzerlik Ölçüsü

Öklidyen olmayan bir benzerlik ölçüsü geliştirilirken dokümanların konumları hesaplamalarda dikkate alınmalıdır. Kullanılan uzay ve geometrik yapı varsayımına göre dokümanlar belli konumlara eşleştirilir. Spesifik olarak konumları özelleştirmek için veri seti bir veri yapısına eşleştirilebilir (mapping) ya da belli bir geometrik yapıya gömülebilir (embedding).

Önerilen yaklaşımda, konum özelleştirmesini sağlamak üzere veri setini oluşturan dokümanlar arasındaki bağların meydana getirdiği ağ yapısı kullanılabilir. Doküman ağları çizge (graph) veri yapısı ile gösterilir, çizge üzerindeki her dokümanın konumu yerel bağlantısallık derecesi (kümelenme katsayısı) açısından özeldir.

Veri setinden veri yapısına dönüşüm yapmak, aslında veri setinden geometrik bir yapıya varma zincirinin ilk halkasıdır. İkinci halka veri yapısını bir geometrik yapı ile eşleştirir. Dolayısıyla bağlı veri setini çizgeye aktarma manifolda geçiş öncesi takip edilen adımdır. Çizge-manifold eşleştirme yöntemlerinin (Belkin and Niyogi, 2008) varlığı düşünüldüğünde doküman ağına ait komşuluk matrisi ya da çizge kullanılarak elde edilecek yerel bağlantısallık dereceleri bir nevi eğrilik değerleridir. Bu gerçek kümelenme katsayısı değerlerinin eğrilik tahminleri olduğunu belirtmesi açısından Lou (2009) tarafından da ifade edilmektedir. Böylece konum özelleştirmesi eğrilik metriğinin değerlendirmeye katılması manasına gelir.

Doküman ağlarını konumları özelleştirmek için kullanırken dikkat edilmesi gereken husus, kullanılacak yöntemin dokümanlar arası bağları ek bir özellik

(feature) olarak değerlendirmeye kattığı ve oluşan ağın taşıdığı yerel ve global niteliklerin araştırılması gerektiğidir.

Dokümanlar arası bağ bilgisi, bilgi modelinin kullanabileceği ek bir özelliktir. Bu bilginin modelin işleyişine dâhil edilmesi diğer özelliklerle birleştirilmesi, birlikte değerlendirilmesi anlamına gelir. Dolayısıyla konumları bu bağ bilgileri ile özelleştirmek aslında birden fazla özelliği birleştirecek bir yöntem önermek demektir.

Jeodezik uzunluk metin tabanlı bağlı veri setlerinde kullanılabilir. Bu veri setlerinden bilgiye ait birden fazla özellik çıkartmak mümkündür ve metrik birden fazla özelliği biraraya getirmemizi sağlar. Veri setinden elde edilebilecek ilk özellik metin içeriği üzerinde sözcük sayma tabanlı terim frekansıdır. Doküman metninden hesaplanan terim frekansları daha önce ifade edildiği üzere idf mekanizması ile normalize edilir. Normalizasyon sonrası tf-idf, sözcükleri temsil eder hale gelir. tf-idf değerleri ile gösterilen doküman vektörleri Cosine benzerlik ölçüsü kullanılarak karşılaştırılır. Jeodezik uzunlukların hesaplanmasında benzerlik ölçümlerine eğrilik tahmini olarak bağ çizgesi tabanlı kümelenme katsayıları dahil edilir. Bir başka deyişle bağ bilgisinden üretilen bir özellik de dikkate alınmış olur.

Jeodezik uzunluk benzerlik ölçüsünün asıl önemli noktası, bağ bilgisi ile metin tabanlı benzerlik ölçülerini matematiksel bir maliyet fonksiyonu kullanarak birleştirebilmesidir.

Benzerlik ölçüleri; belli bir konuya ilişkin dokümanların bulunması, belli bir kategoriye dâhil olanların belirlenmesi gibi sorulara cevap bulmada ölçüt olarak kullanılır ve varolanlar arasında en iyi seçimin yapılması hedeflendiği için genel olarak bir eniyileme problemi üzerinde çalışır. Bağ bilgisine dayalı derecelendirme mekanizmaları ve bağ bilgisini içine alan dil modelleri mevcuttur ancak farklı özellikleri (Cosine, iç-bağlar, PageRank vs.) biraraya getirebilecek eniyileştirilmiş maliyet fonksiyonları çok az sayıdadır ve bulunması kolay değildir. Dolayısıyla, bu gereksinimi karşılayabilecek eniyileştirilmiş maliyet fonksiyonları formüle edilebiliyorsa, cebirsel çözümler olarak tercih edilir.

Analitik çözümlerin varolmadığı bu alan; öğrenme tabanlı yaklaşımların oldukça popüler olduğu bir alandır. Makine öğrenmesi, özellik birleştirmeyi

sağlayan iyi maliyet fonksiyonlarının bulunması hedefine ulaşmak üzere mekanizmalar ortaya koyar.

Bağ bilgisini işleyen yöntemlerde dikkat edilmesi gereken bir başka nokta, doküman bağ ağının niteliğidir. Doküman bağ çizgesinden elde edilen yerel özellikler önemlidir fakat aynı derecede önemli olan, ağda gözlenen yerel özelliklerin global bir sonuç (davranış) ortaya koyup koymadığıdır. Bu genel bakış değerlendirmelerde farklılık yaratabilecek niteliktedir.

Nitekim genel olarak doküman ağları sosyal ağ niteliği taşır ve bu nitelik nedeniyle temsil ettiği çizge ya da komşuluk matrisi özeldir. Sosyal ağ niteliği ile kastedilen, doküman ağının ölçekten bağımsız (scale free) özellikler taşıyıp bir dokümanın diğer k sayıdaki dokümana bağlanmış olma olasılığının power law dağılımı göstermesidir (Barabási et al., 2000).

Power law dağılımı nedeniyle bağlantı yapısı yoğun olan çok az sayıda düğüm vardır, bu düğümler pozitif eğriliğe sahiptir ve ağı birleştirici ve belirleyici nitelik taşımaktadır. Çünkü diğer düğümler aslında bu çekirdek düğümler etrafında yerleşmiştir. Geri kalan, bağlantı yapısı seyrek olan düğümler ise negatif eğrilik değerine sahiptir ve çoğunluğu temsil eder.

Bir doküman ağının sosyal ağ niteliği taşıması yanında hiyerarşik modülerlik (hierarchical modularity) (Ravasz et al., 2002) niteliği gösterip göstermediği bulunabilir. Hiyerarşik modülerlik, ağın sosyal ağ niteliği taşımasına ek olarak, modüler bir yapıda olup olmadığının da bir göstergesidir. Başka bir ifade ile yüksek bağlantı sayısına sahip düğümler ağı birleştirici nitelik taşımaya devam ederken, dokümanlar arasında düzenli kümelenmeler mevcuttur. Bir derlem ile birlikte kategori üst bilgisi verilmekte ise kategori bazında yerel bağlantısallık ortalamalarının değerlendirmesi yapılarak doküman ağının hiyerarşik modülerlik taşıyıp taşımadığı test edilebilir.

Doküman ağlarının oluşturduğu manifold değerlendirilmek istenirse, manifold boyunca görülen sabit eğrilik değerlerinin güçlü global özelliklerin göstergesi olduğu gerçeği ışığında doküman ağlarının hiperbolik bir yapıya sahip olduğu sonucu çıkarılabilir. Bununla birlikte küresel geometrik yapıdaki belirleyici düğümlerin varlığı da dikkate alınmalıdır.

Özetle bağlı bir veri seti için dokümanlar arasındaki bağ bilgisi kullanılarak oluşturulmuş doküman ağına ait bir özellik değerlendirilmek isteniyorsa aşağıda listelenen maddeler dikkate alınmalıdır:

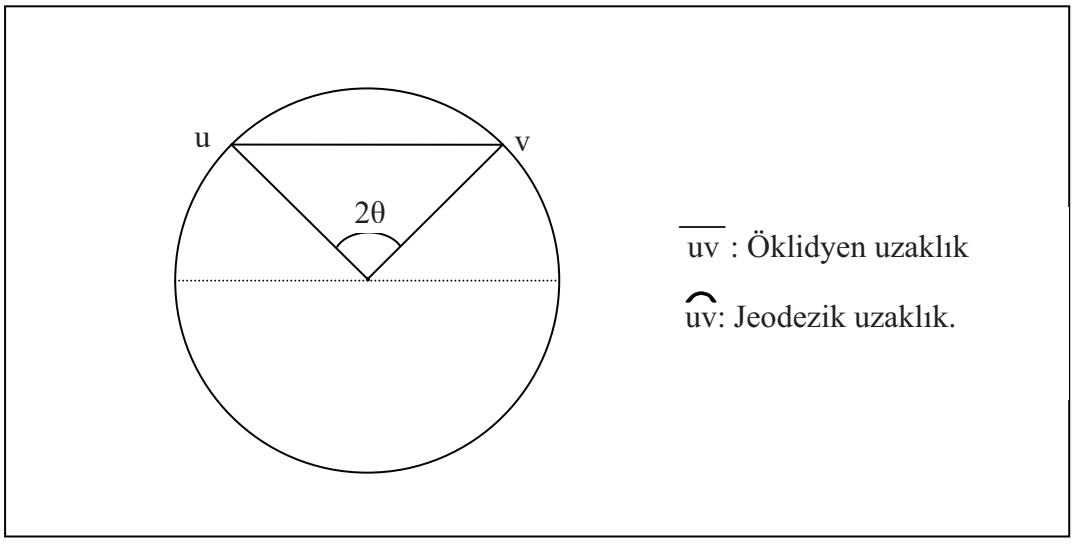
1. Doküman ağlarının özelliklerinin araştırılması, çizge-manifold eşleştirme yöntemlerinin kullanılabilirliği.

2. Eğrilik hesaplama yaklaşımlarının ele alınması.

3. Derleme özgü karakteristiklerin ortaya konması.

3.2 Jeodezik Uzunluk Hesaplama Yöntemi

Bu tez kapsamında, metin tabanlı bağlı veri setleri için Öklidyen olmayan bir benzerlik ölçüsü olarak jeodezik uzunluk önerilmektedir. Önerilen jeodezik uzunluk küresel geometri temel alınarak hesaplanmaktadır. Hesaplama, birim çember üzerinde Öklid ve jeodezik uzaklıklar arasındaki ilişki kullanılarak gerçekleştirilir. Çember üzerindeki iki noktayı birleştiren doğrunun uzunluğu Öklid uzaklığını temsil ederken bu iki nokta arasındaki çember yayı uzunluğu jeodezik ile karşılanır (Şekil 3.1).



Şekil 3.1 Çember üzerinde Öklidyen ve jeodezik uzunluklar.

İki nokta arasındaki doğrusal uzaklık, köşeleri bu iki nokta ve çember merkezi olan üçgen üzerinden hesaplanır ve aşağıdaki şekilde gösterilebilir:

$$d_E(u,v) = 2r\sin\theta \qquad (3.4)$$

İlgili çember yay uzunluğu ise aşağıdaki formül ile verilmektedir:

$$d_g(u,v) = 2r\theta \qquad (3.5)$$

İki nokta arasındaki Öklid uzaklık formülünde yer alan sinθ değeri Maclaurin serisi açılımı ile tahminlenebilir:

$$d_E(u,v) = 2r(\theta - \frac{1}{6}\theta^3 + ...) \qquad (3.6)$$

Jeodezik yay uzunluğu formülünde θ yalnız bırakılarak 3.6 nolu eşitlikte yerine konulursa aşağıdaki eşitlik elde edilir:

$$d_E(u,v) = d_g(u,v) - \frac{d_g^3(u,v)}{24r^2} \qquad (3.7)$$

Son olarak 3.7 nolu eşitlikte çember yarıçapı yerine yarıçapın çember eğriliği cinsinden eşiti konulur;

$$r = \frac{1}{\kappa} \qquad (3.8)$$

ve oluşan eşitlik jeodezik uzunluk bağımlı değişkeni için çözülür.

$$d_g^3(u,v) - 24\frac{1}{\kappa^2}d_g(u,v) + 24\frac{1}{\kappa^2}d_E(u,v) = 0 \qquad (3.9)$$

Denklemdeki bağımsız değişkenler; eğrilik değerleri (κ) ve Öklidyen uzaklıktır (d_E). Tezdeki uygulama açısından, eğrilik değeri yerine ilgili dokümanların kümelenme katsayısı değerleri (veri setine ait komşuluk matrisi kullanılarak) konulmuştur. Öklidyen uzaklık ise metin tabanlı sistemlerde temel benzerlik ölçütü olan Cosine ile değiştirilir.

Denklem 3.6'da verilen eşitlikteki Maclaurin serisi açılımı bir terim daha eklenerek aşağıdaki şekilde genişletilebilir:

$$d_E(u,v) = 2r(\theta - \frac{1}{6}\theta^3 + \frac{1}{120}\theta^5 ...) \qquad (3.10)$$

Oluşan eşitlikte θ yerine jeodezik yay uzunluğu formülündeki eşiti konarak 5. dereceden bir denklem elde edilir.

$$d_E(u,v) = d_g(u,v) - \frac{d_g^3(u,v)}{24r^2} + \frac{d_g^5(u,v)}{1920r^4} \qquad (3.11)$$

Elde edilen denklemde yarıçap değeri eğrilik karşılığı ile değiştirilerek aşağıdaki 5. dereceden eşitlik oluşturulur:

$$d_g^5(u,v) - 80\frac{1}{\kappa^2} d_g{}^3(u,v) + 1920\frac{1}{\kappa^4} d_g(u,v) - 1920\frac{1}{\kappa^4} d_E(u,v) = 0 \qquad (3.12)$$

Kümelenme katsayısı bir çizgede bir düğümün komşuluğundaki bağlantısallık derecesinin ölçüsüdür. Bir düğümün komşuluğundaki düğümler arasında varolan bağ sayısının olabilecek en fazla bağ sayısına oranıdır. Komşularının tamamı ikili olarak birbirine bağlı olan düğümler için kümelenme katsayısı değeri 1 iken komşuları arasında hiç bir bağlantı olmayan düğümler 0 kümelenme katsayısı değerine sahiptir.

Kümelenme katsayıları skaler eğrilik değerlerini tahminlemekte kullanılır. 0.5 değeri 0 eğrilik değerine karşılık gelirken 0.5 ile 1 arasındaki değerler pozitif skaler eğriliği karşılar. Benzer şekilde 0-0.5 aralığındaki değerler negatif skaler eğriliğinin göstergesidir.

Denklem 3.9'da verilen jeodezik uzunluk denklemi ikinci dereceden terimin katsayısının 0 olduğu özel bir üçüncü dereceden denklemdir. Büküm noktasının apsisi ya da köklerin ortalaması sıfırdır (White and Kalman). Üçüncü dereceden bir denklemde büküm noktası fonksiyonun grafiğinde ikinci türevin (iç bükeyliği ya da eğrilik değişimini temsil eder) işaret değiştirdiği noktadır. Büküm noktasının apsisi denklemin köklerinin gerçel kısımlarının ortalamasını verir. Denklemin çözümü için üçüncü derece denklemler için Cardano yöntemi (White and Kalman) uygulanır.

Öte yandan denklem 3.12'de verilen 5. dereceden jeodezik uzunluk denklemi Newton-Raphson nümerik analiz yöntemi ile çözülür.

Bu bölümü izleyen uygulama bölümünde jeodezik benzerlik ölçüsü Wikipedia XML Corpus (Denoyer and Gallinari, 2006) üzerinde kümelenme (clustering)

bağlamında test edilmektedir. Elde edilen duyarlık (precision) değerleri Cosine değerleri ile karşılaştırmalı olarak analiz edilmektedir.

4. UYGULAMA

Jeodezik uzunluklar, içinde bulunulan uzayın doğasına uygun bir şekilde hesaplanır ve uzaydaki eğrilik tanımına bağlıdır. Tezin uygulama kısmında küresel geometri varsayımı temel alınarak hesaplanan jeodezik uzunluklar kullanılmıştır. Hedef, jeodezik uzunlukların doküman uzaylarında semantik benzerlik ölçüsü olarak kullanılabilirliklerini test etmektir.

4.1 Veri Seti ve Ayarlar

Uygulamayı gerçekleştirmek üzere metin tabanlı bağlı bir veri seti olan Wikipedia XML Corpus (Denoyer and Gallinari, 2006) seçilmiştir. Çalışmada bu külliyatın ingilizce alt kümesi kullanılmıştır. Derlem 668058 adet dokümandan oluşmaktadır ve her doküman yarı yapılandırılmış (semi-structured) XML formatındadır. Metin tabanlı analiz işleminde XML dosyalarının metin içeriği alınmış ve çözümleme (parsing) sonucunda çekilen metin, alfabe dışı karakterlerden arındırılmıştır. Metinsel analizde XML etiketleri (tags) dikkate alınmamıştır. Metin indeksleme ve külliyat sözlüğü oluşturma işlemleri Sujit Pal tarafından yazılmış "Java Text Mining Tools" (Pal, 2009) yazılım aracı kullanılarak gerçekleştirilmiştir. Sözcük torbası (Bag-of-words) oluşturulurken doküman metnine sözcüklerin sınırlarını kontrol etmek üzere sınır tanıyıcı (boundary recognizer), her dokümanda ortak kullanılıp dokümana özgü olmayan sözcüklerin değerlendirilmesi için atlanacak kelime (stopword) tanıyıcı ve içeriği tanımlayıcı kelimeleri tayin etmek üzere sözlük veritabanı Wordnet 2.1'i (Miller, 2009) kullanan içerik sözcüğü tanıyıcı (content word recognizer) uygulanmaktadır.

Dokümanlar arası bağlara ait komşuluk matrisini oluşturmak amacı ile derleme ait dokümanlar <collectionlink> etiketi açısından çözümlenerek bağdan bağa veritabanı (link-to-link database) oluşturulmuştur. <collectionlink> etiketi derlem içerisindeki dokümanların birbirlerine verdikleri referansları tutar. Bu çalışma kapsamında derlem dışına çıkan ya da doğrudan bağımsız bir Wikipedia dokümanını göstermeyen bağlar ihmal edilmiştir.

Bağdan bağa veritabanı oluşturulduktan sonra bu veritabanı bir komşuluk matrisine aktarılır. Bu aktarım işleminde dikkat edilecek iki husus vardır:

1. Komşuluk matrisi çok büyük seyrek (sparse) bir matristir.

Çok büyük matrislerin bellekte tutulması ve işlenmesi oldukça maliyetlidir. Elde edilen komşuluk matrisi çok büyük olmasının yanında seyrek de olduğu için veriyi daha "compact" bir formda tutmak mümkündür. Bu amaçla seyrek matrislerin etkin bir şekilde tutulduğu veri yapılarını destekleyen kütüphane ya da yazılım ortamlarına ihtiyaç vardır.

2. Bağdan bağa dosyası işlenirken bir etiketleme algoritması oluşturulması gerekir.

Matris elemanlarına erişim, indisler üzerinden gerçekleşir ve indisler en küçük değerden başlayarak sıra ile artırılır. Etiketleme işlemi de okunan ilk doküman ID'sine en küçük etiketi vererek sırasıyla artan etiketler oluşturur. Bu sayede matris üzerinde ardışık indisler yolu ile hareket edilir, bellek verimli bir şekilde kullanılır.

Etiketleme mekanizmasını seyrek matris kavramı ile birleştiren bir algoritma yardımıyla dokümanlar arasındaki bağ bilgileri seyrek (sparse) bir gösterim formatına dönüştürülerek seyrek matris uygulamaları sunan matematiksel yazılımlara otomatik olarak yüklenecek hale getirilir.

Derlemdeki her bir doküman için, derlemin tamamına ait bağ bilgisinden oluşturulan derlem komşuluk matrisi üzerinde kümelenme katsayısı hesabı yapılır. Derlem komşuluk matrisi ve etiketler vektörünü parametre olarak alıp kümelenme katsayısı vektörünü oluşturan kümelenme katsayısı prosedürünün prototipi, Algoritma 4.1'de verilmektedir:

Algoritma 4.1 Kümelenme Katsayısı

```
procedure clusteringcoef
     (adj: komşuluk matrisi, labels: etiketler vektörü)
returns ccoef[i].
```

Wikipedia XML Corpus'a ait iç-bağ (in-link), dış-bağ (out-link) ve kümelenme katsayısı istatistikleri ise Çizelge 4.1'de görülmektedir:

Çizelge 4.1 Wikipedia XML Corpus iç-bağ, dış-bağ ve kümelenme katsayısı istatistikleri.

	min	**max**	**ortalama**	**medyan**	**standart sapma**
iç-bağ	0	74950	20,9016	4	289,0161
dış-bağ	0	5176	20,9016	12	37,3416
kümelenme katsayısı	0	1	0,2493	0,2	0,1875

Bir niceliğin belli bir değerinin görülme olasılığı o değerin bir kuvveti ile ters orantılı ise niceliğin power-law dağılışı gösterdiği söylenir ve bu yasa Zipf yasası olarak da geçmektedir (Newman, 2005). Matematiksel olarak power-law olasılık dağılışı $P(x) \sim x^{-\gamma}$ fonksiyonu ile ifade edilmektedir. Bu fonksiyonun üssü γ, dağılışın sabit parametresidir ve bazı istisnai durumlar dışında 2 ile 3 arasında değerler alır. Pratikte çok az sayıdaki olgu, olası tüm x değerleri için power-law dağılışı gösterir. Genelde, power-law belli bir sayıdan büyük x değerleri üzerinde geçerli olmaktadır. Dolayısıyla power-law testlerinde bu minimum x değeri, dağılışın sabit parametresi ile birlikte hesaplanmaktadır. Wikipedia XML Corpus iç-bağ ve dış-bağ dağılışlarının power-law özelliği gösterip göstermediğini test etmek üzere Clauset et al. (2009) tarafından önerilen aşağıdaki prosedür takip edilmiştir:

- Power-law modelinin x_{min} ve sabit γ parametrelerinin tahminlenmesi.

- Veri ve power-law uyum iyiliği (goodness-of-fit) testinin yapılarak ilgili p değerinin hesaplanması. Hesaplanan p değerinin 0.1'den büyük olması durumunda power-law'un veriyi temsil etmesi makul bir hipotez olarak kabul edilirken 0.1'den küçük p değerleri için hipotezin reddedilmesi.

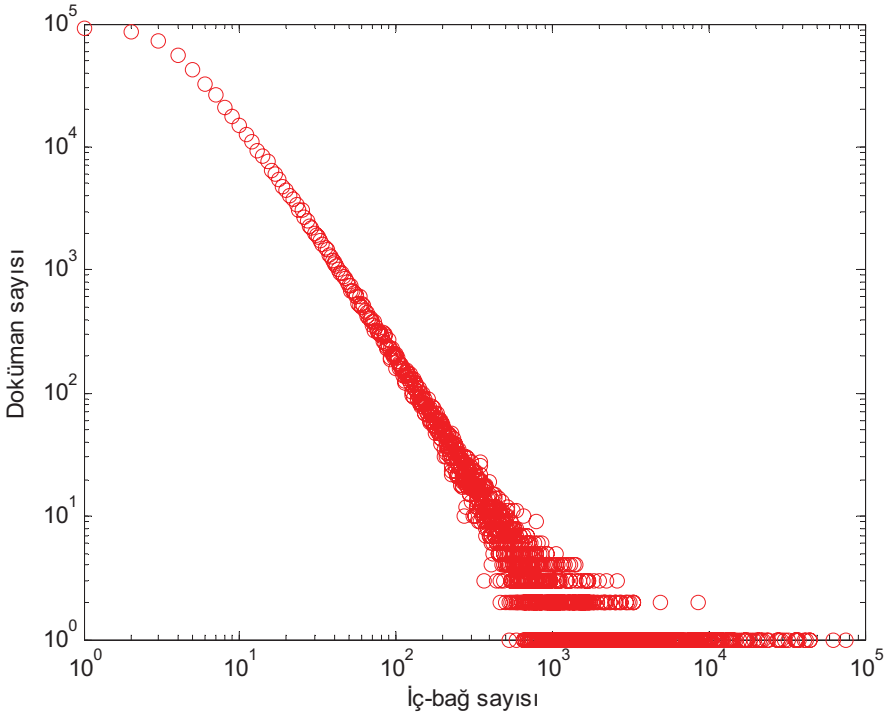
- Power-law'a benzer nitelikler taşıyan üssel dağılış gibi diğer dağılışlar için de benzerlik oranı testlerinin yapılarak bu testler sonucunda elde edilen p değerlerinin power-law durumunda hesaplanan p değeri ile karşılaştırılması suretiyle büyük olan p değerine ait dağılışın kabul edilmesi.

Gerçekleştirilen testler sonucunda iç-bağlar için power-law sabit parametresi $\gamma = 2.21$, x_{min} değeri ise $x_{min} = 280$ olarak bulunmuştur. Uyum iyiliğini gösteren p değeri ise $p = 0.1429$ olarak hesaplanmıştır. Sonuç olarak 0.1'den büyük olan p, dağılışın istatistiksel olarak anlamlı olmak üzere power-law dağılışı gösterdiğini ifade etmektedir. İç-bağlar üzerinde üssel dağılış için yapılan benzerlik oranı testleri anlamlı bir p değeri döndürmemiştir.

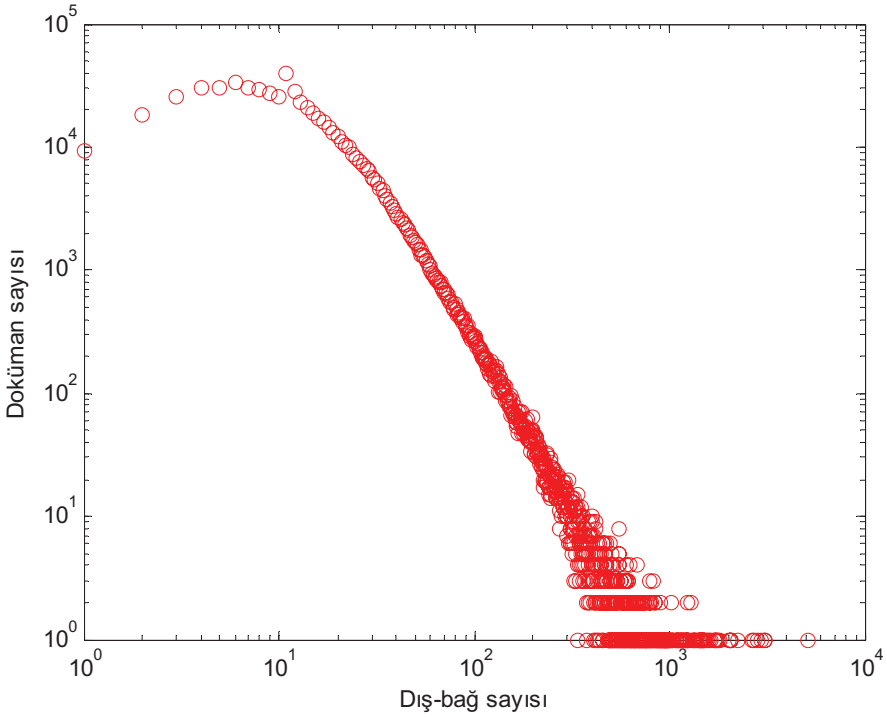
Aynı denemeler dış-bağlar için yapıldığında sırasıyla $\gamma = 2.75$ ve $x_{min} = 39$ değerleri elde edilmiştir. Prosedürün ikinci aşamasında hesaplanan p değeri 0'a çok yakın olduğu için power-law dağılışı dış-bağ verisini istatistiksel anlam sınırları içerisinde temsil etmemektedir. Aynı şekilde üssel dağılış testinde de anlamlı bir p değeri elde edilememiştir.

İç-bağ power-law dağılışı Şekil 4.1'deki grafikte de gözlenmektedir. Dış-bağlara ait olan Şekil 4.2'deki grafik ise iç bağlardaki kadar belirgin olmasa dahi benzer bir eğilimi göstermektedir. Şekil 4.1 ve 4.2 sırasıyla iç-bağ ve dış-bağ histogramlarının logaritmik ölçekte verilmiş halidir ve logaritmik ölçekte doğrusal bir grafikle karşılaşılması ve oluşan doğrusal yapının eğiminin negatif olması power-law olasılık dağılışı formülü ($P(x) \sim x^{-\gamma}$) gözönünde bulundurulduğunda dağılışın power-law'a yakın olduğunun göstergesidir.

Wikipedia iç ve dış-bağ dağılış grafikleri birbirine benzerlik gösterir. Web'de ise iç-bağlarda power-law dağılışı görülürken dış-bağların iç-bağ davranışından oldukça farklılaştığı gözlenmektedir. Dolayısıyla bağ davranışları açısından Wikipedia ve Web farklıdır. Bu farklılık Wikipedia'nın nispeten kontrollü bir ortam olmasından kaynaklanmaktadır. Ayrıca iç-bağ bilgisi dokümanların anlamsal benzerliklerini ölçmede Web'e göre Wikipedia'da daha etkilidir ve yine Wikipedia'da dış-bağ bilgisi de anlamsal yakınlıklar açısından önem taşımaktadır (Kamps and Koolen, 2009).



Şekil 4.1 Wikipedia XML Corpus iç-bağ power-law dağılışı.



Şekil 4.2 Wikipedia XML Corpus dış-bağ power-law dağılışı.

Çizelge 4.1'den hareketle (ortalama ve medyan dikkate alınarak) Wikipedia XML Corpus için kümelenme katsayısı değerlerinin 0.5'ten küçük olma eğiliminde olduğu görülür ve bu değerlerin eğrilik tahminleri olarak kullanılabileceği değerlendirildiğinde doküman uzayında eğriliğin genel olarak negatif olduğu sonucuna varılabilir.

Uygulanan yaklaşımda bağ bulgusu (link evidence) olarak kümelenme katsayısı değerleri kullanılır. Kümelenme katsayısı hesabı yönsüz bir çizge üzerinde yapıldığından ve Wikipedia'da bağların dokümanların anlamsal benzerlikleri üzerinde simetrik bir etkisi olduğundan (A ve B arasındaki anlamsal benzerlik ilişkisi hem iç hem de dış bağlar yoluyla ifade edilir. Bir başka deyişle A, B'ye anlamsal olarak yakınsa B de A ile benzer anlamdadır) (Kamps and Koolen, 2009) bu yaklaşım özellikle Wikipedia için geçerlidir.

Wikipedia XML Corpus içerisinde yer alan her Wikipedia sayfası için, ait olduğu kategori bilgisi verilmektedir. Bu üst bilgi (metadata) iki dosya halinde sunulmaktadır. Kategoriler dosyası kategori adları ve bu adlara karşılık gelen ID'leri içerir. Kategori-doküman eşleştirmelerini tutan ikinci dosya ise doküman ID, kategori ID ikililerinden oluşmaktadır. Derlemde toplam 72 ingilizce Wikipedia portal kategorisi bulunmaktadır.

Derlem ile birlikte gelen kategori bilgileri, semantik benzerlik ölçüsü olarak jeodezik uzunlukların işe yararlığını test etmek üzere kümelenme (clustering) uygulamalarını uygun seçenekler haline getirir. Jeodezik uzunluk, kümelenme bakış açısı ile değerlendirildiğinde aynı kategoriye ait olan dokümanların metin benzerliklerinin yüksek olduğu ve global bağ çizgesi baz alınarak hesaplanan kümelenme katsayısı değerlerinin metin benzerlik ölçülerine faydalı ağırlıklar sağlayacağı ifade edilebilir.

Tüm kategoriler üzerinde ayrıntılı testlerin yapılması, zaman ve bellek açısından oldukça maliyetlidir bu nedenle sınamalar için veri setinden rastgele olarak 10 Wikipedia kategorisi seçilmiştir. Çizelge 4.2'de, seçilen kategoriler ve bu kategorilere ait doküman sayıları verilmektedir:

Çizelge 4.2 Seçilmiş kategoriler ve büyüklükleri.

Kategori Adı	**Büyüklük (Doküman Sayısı)**
Venezuela	569
Bangladesh	393
Colombia	304
Hong Kong	11056
Romania	1340
New Zealand	2393
Morocco	230
Finland	1887
Netherlands	1350
Uganda	232
Toplam	19754

Dokümanlara ait kümelenme katsayıları, veri setine ait global bağ çizgesi üzerinden hesaplanır çünkü düğüm sayısı oldukça arttığında kümelenme katsayısı değerleri belli değerlere yakınsamaya başlar. Bir başka deyişle kararlı hale gelir. Global bağ çizgesi yerine, seçilmiş kategorilere ait çizge üzerinde hesaplama yapılmak istendiğinde dokümanlar arasındaki seyrek bağlar nedeniyle çok sayıdaki düğüm için kümelenme katsayısı değerlerinin hesaplanamadığı görülür. Bir düğümün komşuluğunda ikiden az bağ varsa ikili kombinasyon alma işlemi tanımlanamayacağından kümelenme katsayısı tanımsız hale gelir.

Uygulamada sözcük torbası (bag-of-words) yaklaşımı ile indekslenen her doküman için çok boyutlu (her boyut bir sözcük olmak üzere) bir vektör oluşturulur. Metin tabanlı analiz sonrasında ulaşılan çok boyutlu veri, çoğu kümelenme algoritması için uygun değildir. Zira çok sayıdaki algoritma yüksek boyutlu uzaylarda oldukça zayıf kalmaktadır (curse of dimensionality). Dolayısıyla ilgili kümelenme algoritmalarının seçiminde metin bağlamına uygunluk ön planda olmalıdır.

Strehl et al. (2000) web sayfalarının kümelenmesinde benzerlik ölçülerinin değerlendirimi için bir çerçeve sunar. Çalışmada temel benzerlik ölçüleri geometrik yorumları ile tartışılmaktadır. Metin bağlamına uygun kümelenme algoritmaları incelenip mevcut benzerlik ölçüleri ile algoritma performansları karşılaştırılmaktadır. Kullanılacak kümelenme algoritmasının seçiminde bu çalışmadan yararlanılmıştır.

Sınamalarda kümelenme algoritması olarak k-means (MacQueen, 1967) seçilmiştir. k-means algoritması metin tabanlı kümelenme işlerinde yaygın olarak kullanılan popüler bir algoritmadır ve temel parametresi oluşturulacak küme sayısıdır. Bu özelliği, içinde bulunulan duruma uygundur çünkü veri seti dokümanlarına ilişkin kategori sayısı önceden bilinmektedir. k-means'in isabetli bir tercih olmasını sağlayan bir başka nokta ise bu algoritma için birden fazla özellik uzayını birleştirme olanağı sunan soyut bir çerçevenin ortaya konmuş olmasıdır (Modha and Spangler, 2003). Bu sayede çalışma kapsamında varolan iki ayrı özellik kümesinin (eğrilik değerleri ve sözcük vektörleri) birlikte değerlendirilmesi mümkün hale gelir.

k-means algoritması, oluşturulacak küme sayısı k'yı parametre olarak alır ve aşağıdaki adımları işletir:

1. k tane farklı altküme oluşturulur,

2. her altküme için altküme merkez noktası (centroid) hesaplanır,

3. yerleştirilecek nesneler altküme merkezlerine yakınlıkları açısından değerlendirilip en yakın oldukları merkezin altkümesine yerleştirilir,

4. 2. adım yeni bir değişiklik olmayıncaya kadar yinelenir.

Kümelenme algoritmalarının hedefi çarpıklık (distortion) ölçüleri üzerinden tanımlanır. Çarpıklık ölçüleri algoritma sonucu elde edilen kümelerle kesin referans (ground truth) kategorizasyon arasındaki sapmayı ölçer. Dolayısıyla ilgili sapmayı kontrol altında tutmak gerekir. Çarpıklık ölçüleri aynı kategorideki dokümanlar arasındaki uzaklığı azaltacak şekilde çalışmalıdır.

k-means algoritmasında bunun karşılığı doküman-küme merkezi uzaklıklarını küçük tutmaktır. Bir başka ifade ile algoritma kapsamında ayrık (disjoint) k adet

küme, algoritmaya ait maliyet fonksiyonunu (objective function) minimize edecek şekilde oluşturulmalıdır. k-means algoritmasına ait maliyet fonksiyonu aşağıda verilmektedir (Modha and Spangler, 2003):

$$\{\pi_u\}_{u=1}^{k} = \arg\min_{\{\pi_u\}_{u=1}^{k}} \left(\sum_{u=1}^{k} \sum_{x \in \pi_u} D^{\alpha}(x, c_u) \right) \qquad (4.1)$$

Verilen denklemde c_u küme merkezi vektörünü belirtir.

Ağırlıklandırılmış çarpıklık (weighted distortion) ölçüleri aşağıdaki formül ile tanımlanabilir (Modha and Spangler, 2003):

$$D^{\alpha}(x, x') = \sum_{l=1}^{m} \alpha_l D_l\left(F_l, F_l^{'}\right) \qquad (4.2)$$

Formülde D çarpıklık ölçüsü, α ağırlık, x oluşan kümelerin kümesi, x' kategori kümesi ve F_l ve $F_l^{'}$ sırasıyla kümeler ve kategoriler kümesindeki l. özellik (feature) vektörleridir. Ağırlık vektörleri veri setinden elde edilen özellikler üzerinde ağırlıklandırma görevi üstlenir.

k-means algoritmasında önerilen şekli ile jeodezik uzunluk ağırlıklandırılmış çarpıklık ölçüsü özelliği gösterir. Tanımlanan bağlamda Cosine esas çarpıklık ölçüsü iken veri setine ait bağ çizgesinden gelen ağırlıklar Cosine ile birlikte ağırlıklandırılmış çarpıklık ölçüsünü meydana getirir.

Sınamalarda Cosine benzerlik ölçüsü ile birlikte k-means, kontrol değişkeni kabul edilirken k-means algoritmasına jeodezik uzunluk kavramı uyarlanmıştır. İlk olarak veri setinin global komşuluk matrisinden hesaplanan kümelenme katsayısı değerleri her doküman için kaydedilir. k-means algoritması her küme için o kümeye dahil edilen doküman vektörlerinin ortalaması ile temsil edilen bir kümelenme merkezi (centroid) kavramı çerçevesinde geliştirilmiştir. Her doküman için hesaplanan kümelenme katsayısı değerinin kümelenme merkezi kavramına uyarlanması gerekmektedir. Metin tabanlı kullanımına benzer şekilde, kümelenme merkezinin eğriliğini karşılamak üzere o kümeye atanmış dokümanların kümelenme katsayılarının ortalaması alınır. Küme merkezi vektörü ve eğrilik değeri aşağıdaki şekilde özetlenebilir:

Küme merkezi vektörü (centroid) μ_c: C kümesine atanmış doküman vektörlerinin ortalaması.

Küme merkezi eğrilik değeri κ_c: C kümesine atanmış dokümanların eğrilik değeri ortalaması.

Kümelenme katsayısı değerlerinin eğrilik değerlerinin kaba bir tahmini olduğu varsayılarak (Lou, 2009) bu ortalama kümelenme katsayılarının ayırt ediciliğini artırmak üzere basit bir sezgisel algoritma (heuristic) uygulanır. Kümelenme katsayısı değerlerinden eğrilik değerlerini oluşturmak için kullanılan sezgisel algoritma kodu Algoritma 4.2'de verilmektedir:

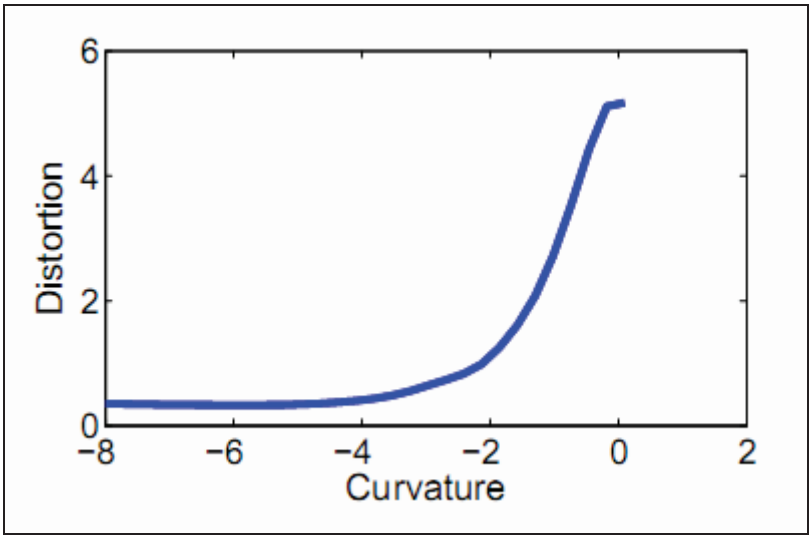
Algoritma 4.2 Kümelenme katsayısı değerlerini eğrilik değerlerine dönüştürmek için kullanılan sezgisel algoritma.

```
input: ccoef   // kümelenme katsayısı
test koşulu: eğrilik değerinin pozitif olup olmadığı;
if ccoef>0.5 then
   ccoef=ccoef-0.5
else
   ccoef=ccoef+1
end
```

Veri setindeki dokümanların kümelenme katsayısı değerleri negatif olma eğilimindedir ve ortalaması 0.2'dir (Bkz. Çizelge 4.1). Bu çok sayıdaki negatif eğrilik değeri dokümanların hiperbolik bir uzayda yerleştiğini gösterir. Algoritma 4.2'de tanımlanan sezgisel algoritmada 0.5 kümelenme katsayısı değerine yakın olan negatif eğrilik değerlerine sahip dokümanlar için geçerli olan çarpıklık (distortion), 0.5 değerinden uzak olan negatif eğrilik değerine sahip dokümanlara göre daha fazladır. Bir başka deyişle negatif eğrilik değerlerinin kendi içerisindeki sıralaması her bir değere 1 eklenmek suretiyle değiştirilir.

Bu çarpıklık varsayımı Internet'in hiperbolik bir uzaya gömüldüğü (embedding) Begelfor et al. (2005) deneysel sonuçları ile örtüşmektedir. Nitekim

gömülme işleminde ortaya çıkan çarpıklık değerlerinin negatif eğrilik değerlerine bağlı grafikleri sezgisel algoritmadaki negatif eğrilik-çarpıklık ilişkisini doğrular niteliktedir (Şekil 4.3). Ayrıca sezgisel algoritma içerisinde pozitif eğrilik değerleri arasındaki sıralama korunur, bu değerlerin kümelenme merkezinin eğriliğine etkisi yapılan çıkarma işlemi kanalıyla azaltılır.



Şekil 4.3 Internetin iki boyuta gömüldüğü (embedding) durumda görülen çarpıklığın (distortion) gömülme uzayının eğriliğinin bir fonksiyonu olarak gösterimi (Negatif değerler hiperbolik, pozitif değerler ise küresel geometriyi temsil etmektedir) (Begelfor et al., 2005).

Sözü geçen sezgisel algoritmanın uygulanmasının ardından aynı küme merkezi etrafında kümelenmiş dokümanların eğrilik değerlerinin aritmetik ortalaması hesaplanarak küme merkezi eğrilik değeri elde edilir. Bu hesaplama işleminde az sayıdaki, kümelenme katsayısı değeri tanımlı olmayan (NaN) doküman ihmal edilmiştir. İhmal edilen bu dokümanlar komşuluklarında birden fazla bağ olmadığı için eğrilik tahmini yapılamayan düğümlerdir. Çünkü kümelenme katsayısı yaklaşımı ile eğrilik hesabı ikili kombinasyon alma işlemine bağlıdır ve ikiden küçük sayılar için ikili kombinasyon tanımlı değildir.

Eğriliklerin hesaplanmasının ardından çember üzerindeki jeodezik uzunluk formülü uygulanır (3.7). Formülün uygulanması ile elde edilen 3. dereceden denklemde ilgili katsayılar yerine konarak denklemin kökleri hesaplanır. Daha önceden de belirtildiği gibi denklemin reel kökü jeodezik uzunluk olarak alınır. Aşağıda denklemin çözümüne ilişkin kod satırları görülmektedir:

Algoritma 4.3 k-means algoritmasına uyarlanan jeodezik uzunluk benzerlik ölçüsünün hesaplanması.

```
Complex[] roots_1 =
GeodesicSimilarity.cubicfcn
(1,0,-24/Math.pow(centroid_ccoef,2),24*similarity);


similarity=roots_1[1].abs();
```

Verilen kod parçasındaki `roots_1` dizisi 3. dereceden denklemin köklerini tutmaktadır. `centroid_ccoef` dokümanın atandığı kümenin küme merkezi eğriliğini temsil etmekte, fonksiyona parametre olarak geçirilen `similarity` ise doküman ile atandığı küme merkezi arasındaki Cosine benzerlik ölçüsü değerini içermektedir. Son satırda `similarity` değişkenine 3. dereceden denklemin reel kökü olan jeodezik benzerlik ölçüsü değeri atanmaktadır.

4.2 Karşılaştırmalı Değerlendirme

Karşılaştırmalı analiz yaparken başlangıç koşullarını sabit tutmak gerekir. Bu bağlamda Cosine ve jeodezik benzerlik ölçülerini adil (fair) bir şekilde karşılaştırmak için k-means algoritmasının başlangıç küme merkezleri sabitlenir. Cosine durumunda küme merkezleri rastgele olarak belirlenir. Sonrasında Cosine durumunda belirlenmiş olan başlangıç küme merkezleri jeodezik denemesi küme merkezleri olarak elle atanır. Kısacası her denemede Cosine ve jeodezik aynı başlangıç ayarları ile çalıştırılır. Bununla birlikte Cosine-jeodezik ikilisinden oluşan her denemede geçerli olan başlangıç küme merkezleri birbirinden farklıdır.

Yukarıda açıklandığı üzere; 10 deneme gerçekleştirilmiştir ve denemelerde algoritmanın temel parametresi olan oluşturulacak küme sayısı, seçilmiş kategori sayısının iki katı olarak belirlenmiştir. Küme sayısını kategori sayısına eşitlemek yerine kategori sayısının iki katı olarak atamak benzerlik ölçülerinin etkilerinin daha açık bir şekilde gözlemlenmesini sağlar. Strehl et al. ‘da (2000) oluşturulacak küme sayısı bilinen kategori sayısının iki katı olarak belirlenmiş ve bu ayarın başlangıç denemeleri ve görselleştirme (visualization) sonuçları tarafından doğrulanan daha uygun kümeleri oluşturduğu ortaya konmuştur.

Kümelenme sonuçları; ikili sayma (pair counting) tabanlı rand indisi (Rand, 1971) ve ayarlanmış (adjusted) rand indisi (AR) ve bilgi kuramına dayanan ortak bilgi (mutual information) (Cover and Thomas, 2006; Strehl et al., 2000) ile normalleştirilmiş ortak bilgi (NMI) metrikleri yardımı ile değerlendirilir. Wu et al (Wu et al., 2009) tarafından k-means algoritması için en uygun metriklerin van Dongen, rand indisi ve ortak bilgi metriklerinin normalleştirilmiş halleri olduğu gösterilmiştir.

Bu metriklerin hesaplanması için kategori etiketleri vektörü ile küme etiketleri vektörüne gereksinim duyulur. Algoritma 4.3 bu metriklerin hesaplanması için kullanılan fonksiyonların ortak prototipini vermektedir:

Algoritma 4.4 Kümelenme Metrik Hesabı Fonksiyonlarının Ortak Prototipi.

```
procedure calculateMetric
     (c1: kategori etiketleri vektörü, c2: küme etiketleri
   vektörü)
returns metric value.
```

Kümelenme algoritmaları sonuçlarının dökümü çapraz çizelgelere (contingency table) yapılır. Çapraz çizelgelerde, satırlar veri setine ait kategorileri belirtirken sütunlar algoritmanın çalışması sonucunda elde edilen kümeleri temsil eder. Çizelge 4.3'te kategori ve küme kümelerini karşılaştırmayı sağlayan çapraz çizelge gösterimi verilmektedir:

Çizelge 4.3 İki kümeyi karşılaştırmaya yarayan çapraz çizelge gösterimi.

Kategori\Küme	v_1	v_2	...	v_c	Toplam
u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Toplam	$n_{.1}$	$n_{.2}$	...	$n_{.C}$	$n_{..}=n$

Çapraz çizelgeler de, kategori etiketleri vektörü ile küme etiketleri vektörü kullanılarak oluşturulur.

Uygulanan yaklaşımda, k-means algoritmasının temel parametresi olan oluşturulacak küme sayısının bilinen kategori sayısının iki katı olarak atanmasının doğal sonucu olarak bu iki vektörün etiket aralığı birbirinden farklı olmuştur. Kategori etiketleri 1 ile n arasında değişirken, küme etiketleri 0-2n aralığında değer almaktadır. Bir başka ifade ile çapraz çizelge (contingency table) satırlar kategorileri, sütunlar kümeleri temsil etmek üzere n satır ve 2n sütundan oluşmaktadır. Bu durum kümelenme metriklerinin hesaplanmasını zorlaştırır. Güçlük 2n küme arasından kategoriler ile eşleştirilecek n kümenin seçimindedir. Kategori ve küme etiket değerlerinin farklı aralıklarda bulunduğu ihmal edilip metrik hesabı yapıldığında hem Cosine hem de jeodezik uygulamasında kümelenme kalitesinin çok düşük olduğu görülür. Çapraz çizelgeden kategoriler ile kesişim kümesi en büyük olan sütunlar (kümeler) seçildiği takdirde yapılan karşılaştırmaların çok güvenilir olmayacağı ortaya çıkar. Zira bir durumda seçilen en büyük küme ile bir sonrakinin büyüklüğü oldukça yakın olabilirken diğer durumlarda aralarındaki fark çok büyük olabilir.

Uygulama kısmı için seçilmiş kategoriler gözönüne alındığında birden fazla kategoriye ait olan çok az sayıda doküman bulunmaktadır. Başka bir deyişle kategori çakışması ihmal edilebilecek düzeydedir.

Karşılaştırma sonuçlarının değerlendirmesinde kesin referans (ground truth) sınıflandırma ile kümeler arasındaki örtüşmeyi ölçmek üzere duyarlık (precision) sayıları hesaplanmıştır. Veri seti ile birlikte gelen Wikipedia kategorileri kesin referans sınıflandırmayı temsil eder. Denemelerde Wikipedia kategorileri-Cosine benzerlik ölçüsü ile k-means ve kategoriler-jeodezik benzerlik ölçüsü ile k-means kümelenme sonuçları karşılaştırmalı olarak analiz edilmiştir. Duyarlık hesaplamaları Modha and Spangler (2003) tarafından sağlanan yönteme dayalı olarak yapılmıştır. Modha and Spangler (2003) k-means algoritması için birden fazla özellik uzayını birleştirmek üzere kullanılabilecek bir çerçeve tanımlar. Bu çerçeve bu tez çalışması kapsamında yapılan uygulamanın da bağlamını ortaya koyar. k-means algoritması bağlamında tek özellik uzayı ve birden fazla özellik uzayları arasında geçerli karşılaştırmalar yapılması işi de sağlanan çerçevenin sınırları içerisinde kalır. Dolayısıyla sonuçların karşılaştırılmasında geleneksel k-means kümelenme metrikleri NMI ve AR yerine bu çerçeve kapsamında tanımlanan duyarlık metriklerini kullanmak daha makul olur. Denemelerde

duyarlık metriğine ek olarak NMI ve AR değerleri de hesaplanmıştır. NMI ve AR değerlerinin hesaplanmasında Vinh et al. (2009) tarafından sağlanan kod kullanılmıştır. Deneysel sonuçlar; her denemede hem Cosine hem de jeodezik benzerlik ölçüsü için elde edilen NMI ve AR değerleri arasındaki sıralamanın aynı deneme için ölçülen duyarlık değerleri sıralaması ile örtüştüğü görülür. Özetle metrik yönelimlerinin benzerlik taşıdığı, birbiri ile uyuştuğu görülür.

Duyarlık metriğini anlamlı bir şekilde tanımlamak için oluşturulan kümelenmeler aşağıdaki kural izlenerek son sınıflandırmaya dönüştürülür:

- Her kümeyi kesişimi en fazla olan kategori ile ilişkilendir ve o kümedeki her dokümanı aynı kategoriye eşleştir.

Bu kural ile birden fazla küme bir kategoriye atanırken bir küme birden fazla kategori ile eşleştirilemez.

Kesin referans sınıflandırmada n tane kategori $\{\omega_i\}_{i=1}^{n} = 1$ olduğu varsayılırsa duyarlık aşağıdaki eşitlikle ifade edilir:

$$p_i = \frac{a_i}{a_i + b_i}, \quad 1 \le i \le n \qquad (4.3)$$

Eşitlikte a_i, ω_i kategorisine doğru olarak atanmış doküman sayısını, b_i ise ω_i kategorisine yanlışlıkla atanmış doküman sayısını ifade eder.

Duyarlık, kategori bazında tanımlanır. Seçilmiş tüm kategorileri temsil etmek üzere mikro-duyarlık (micro-precision) metriği kullanılır ve her bir duyarlık metriğinin tüm kategoriler için ortalaması alınarak hesaplanır:

$$micro - p = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad (4.4)$$

İlk olarak Çizelge 4.4'te 1. deneme için Cosine benzerlik ölçüsü ile birlikte k-means algoritması sonuçları çapraz çizelgesi verilmektedir. Çizelge 4.5'te ise aynı deneme için jeodezik benzerlik ölçüsü ile k-means algoritması sonuçları gösterilmektedir.

Çizelge 4.4 1. deneme için Cosine benzerlik ölçüsü ile birlikte k-means sonuçları çapraz çizelgesi.

1_cosine_k-means																					
Kategori\Küme	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Toplam
hong_kong	89	425	475	163	522	79	23	636	1070	26	430	487	2989	177	216	664	287	940	1231	127	11056
new_zealand	63	864	33	1084	103	3	7	51	31	1	18	20	28	16	7	7	10	30	16	1	2393
finland	20	30	42	104	25	6	478	97	58	735	36	19	58	21	14	14	21	74	3	32	1887
netherlands	5	163	136	9	25	68	4	261	144	128	16	39	18	28	33	7	9	253	1	3	1350
romania	21	28	57	3	32	639	6	88	155	36	52	42	44	12	11	11	59	39	1	4	1340
venezuela	4	17	10	2	2	1	1	17	4	12	3	82	2	407	1	0	0	4	0	0	569
bangladesh	4	21	12	9	4	3	8	11	10	0	260	5	10	6	3	0	3	22	1	1	393
colombia	3	21	46	3	0	0	0	23	7	2	3	17	12	147	4	0	3	12	1	0	304
uganda	1	7	13	1	11	3	10	3	1	0	165	7	2	2	1	0	4	0	0	1	232
morocco	33	17	24	5	11	8	21	5	28	0	4	4	16	22	0	3	13	14	0	2	230
Toplam	243	1593	848	1383	735	810	558	1192	1508	940	987	722	3179	838	290	706	409	1388	1254	171	19754

Çizelge 4.5 1. deneme için jeodezik benzerlik ölçüsü ile birlikte k-means sonuçları çapraz çizelgesi.

1_geodesic_k-means																					
Kategori\Küme	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Toplam
hong_kong	71	263	442	140	526	62	29	776	1473	45	322	587	3201	113	162	771	206	522	1223	122	11056
new_zealand	164	497	27	1292	111	15	9	62	43	2	24	27	43	19	4	7	6	21	19	1	2393
finland	16	11	41	66	14	3	497	102	81	842	23	22	60	18	9	12	7	24	3	36	1887
netherlands	1	86	138	9	17	91	7	250	195	191	9	48	21	25	27	12	7	211	2	3	1350
romania	4	17	51	3	18	570	13	113	240	72	24	59	56	12	9	16	14	42	2	5	1340
venezuela	0	13	11	2	2	2	1	20	4	13	2	111	2	379	1	0	0	6	0	0	569
bangladesh	0	17	14	8	4	2	11	18	16	2	243	8	13	7	2	0	1	24	0	3	393
colombia	0	14	52	3	0	2	1	34	10	4	2	20	18	131	3	0	2	7	1	0	304
uganda	0	6	13	0	10	3	15	9	7	3	152	8	2	2	1	0	0	0	0	1	232
morocco	4	9	25	3	8	10	25	7	59	2	15	6	21	17	1	4	7	5	0	2	230
Toplam	260	933	814	1526	710	760	608	1391	2128	1176	816	896	3437	723	219	822	250	862	1250	173	19754

Deneysel sonuçlar mikro-duyarlık değerleri cinsinden Çizelge 4.6'da verilmektedir. Çizelgedeki birinci sütunda Cosine ile birlikte k-means, ikinci sütunda jeodezik ile birlikte k-means ve son olarak üçüncü sütunda ise bir jeodezik türevi yöntem ile k-means temsil edilmektedir. İki jeodezik yaklaşım arasındaki temel fark küme merkezi ortalama eğrilik değerinin hesaplanmasındadır. Temel jeodezik yaklaşımda ortalama alınırken eğrilik değerleri işaretlerine bakılmaksızın toplanır. Jeodezik türevinde ise pozitif ve negatif eğrilik değerleri kendi içerisinde toplanır, elde edilen iki toplam arasındaki fark aritmetik ortalamada bölüneni meydana getirir.

Çizelge 4.6 k-means algoritmasının Cosine, jeodezik ve jeodezik türevine göre çalıştırılması sonucunda oluşan kümelenmelere ilişkin mikro-duyarlık değerleri.

	Cosine	**Jeodezik**	**Jeodezik türevi**
1	0,727448	0,737775	0,743191
2	0,72355	0,733117	0,738585
3	0,718285	0,724208	0,724866
4	0,738585	0,74552	0,740812
5	0,662752	0,676774	0,684823
6	0,700618	0,706135	0,696973
7	0,702288	0,705528	0,703756
8	0,678394	0,683912	0,685633
9	0,728612	0,728663	0,720462
10	0,724309	0,708565	0,690088

Farklı benzerlik ölçülerinin kümelenme başarımına etkilerini karşılaştırmak üzere parametrik olmayan Friedman testi uygulanmıştır. Friedman testine girdi olarak sunulan matrisin sütunlarında sırasıyla Cosine, jeodezik ve jeodezik türevi bulunmaktadır. Matrisin satırları farklı denemeleri temsil eder. Çizelge 4.6'da sunulan matris için Friedman testinin sonucu:

$p = 0.0608$ 'dir.

Friedman testi matrisin ilk iki sütununa yani Cosine ve jeodezik durumlarına uygulanırsa

$$p = 0.0114$$

elde edilir.

Friedman testi sütun etkilerinin aynı olduğu hipotezine karşı etkilerin birbirinden farklı olduğu görüşünün test edilmesini sağlar. Elde edilen test sonucu ($p = 0.0608 < 0.1$), sütunlardaki üç benzerlik ölçüsünün %90 güven aralığında birbirinden farklı olduğunu gösterir. İlk iki sütun için hesaplanan $p = 0.0114$ değeri ise yaklaşık %99 güvenle kullanılan iki yöntemin farklı olduğunu ortaya koyar. Bir başka deyişle yöntemler kümelenme üzerinde belirtilen güven aralıkları içerisinde farklı bir etkiye sahiptir.

Çizelge 4.6'daki mikro-duyarlık değerleri incelendiğinde, jeodezik yaklaşımların en başarısız sonuçları en son denemede verdikleri görülmektedir. Bu durumun arkasındaki sebebi bulmak üzere Eşitlik 3.6'daki Maclaurin serisi bir terim daha eklenerek, Eşitlik 3.10'daki şekliyle açılır ve sonuç olarak Eşitlik 3.12'deki 5. dereceden denklem elde edilir. Daha iyi bir tahminlemeyi ortaya koyan 5. dereceden denklem nümerik olarak Newton-Raphson metodu ile çözülür. Çözüm sonucunda elde edilen kökün kümelenme sonuçlarına etkisi analiz edildiğinde, bazı spesifik kategorilerde iyileşme yaratırken diğerlerinde performansın düşmesi nedeniyle hemen hemen aynı mikro-duyarlık değerinin hesaplandığı görülür. Diğer denemeler için elde edilen Cosine ve jeodezik çapraz çizelgeleri incelendiğinde ise benzer şekilde jeodezik yaklaşımların getirdiği katkının özellikle aynı spesifik kategorilerde çok büyük iyileştirmeler sonucu olduğu gözlemlenir. Dolayısıyla; jeodezik yaklaşımların başarısı bazı kategori-spesifik faktörlerle yakından ilişkili olmalıdır. Bu durum Yang'ın (2001) da füzyon durumlarında başarıyı ya da başarısızlığı açıklamak üzere kullanılabilecek faktörlerden biri olarak ortaya koyduğu test derlemine özgü karakteristiklerin önemli olduğunu gösterir. Bununla birlikte, ilgili kategori-spesifik özellik açıkça belirlenmelidir.

Tezin son bölümünde, deneysel kısımda elde edilen deneyimler ışığında uygulamaları daha iyi hale getirmek için yapılması gerekenler ortaya konup çalışmanın genel değerlendirmesi sunulacaktır.

5. SONUÇLAR

Bu tez çalışmasında, bilginin modellenmesi üzerine odaklanılmıştır. Bilginin doğasına uygun bir geometrik uzay/yapı arayışı sözkonusudur. Bu arayış ile cevap verilmeye çalışılan gereksinimler aşağıdaki şekilde özetlenebilir:

- Bilgi yığınları içerisinden işe yarar bilginin çekilmesi istenmektedir. İşe yarar bilgi üç niteliği aynı anda karşılar: Konu ile ilgili olma, doğruluk-hassasiyet ve zamanlılık. Bu bilgi nitelikleri kurallara uygun (formal) olarak temsil edildiği takdirde, işe yarar bilgi tanımlanabilir.

- Varolan IR modellerini (vektör uzayı, olasılık, mantıksal vb.) tek bir çerçeve altında toplamak mümkün olabilir (Rijsbergen, 2004).

- Farklı ortamlardan gelen farklı tipteki verileri kapsayacak genel geçer bir çerçeveye ihtiyaç vardır.

- Bilgiye ait üst bilgilerin uygun şekilde değerlendirilmesi gerekir.

- Semantik seviyede çalışacak bir model oluşturulmalıdır.

Bilgi modellemesinde karşılaşılan en temel problem bilginin nesnel ve öznel niteliklerinin birbirinden kolay bir şekilde ayırt edilemeyişidir. Ayrıca öznel nitelikler insanın dahil olduğu süreçler sonucunda ortaya çıkar ve bu alanda geçerli olan matematiksel yöntemler mevcut değildir.

Bilgi niteliklerini bir hiyerarşi dahilinde ele almak gerekir. Hiyerarşinin en üst seviyesinde bilginin değerini ortaya koyan üç temel bilgi niteliği yer alır. En alt seviyede ise doğrudan ölçülebilir özellikler yerleşir. Modelleme işi doğrudan ölçülebilen özellikler ile dolaylı olarak karşılanabilen bilgi nitelikleri arasındaki karmaşık ilişkileri tanımlamaktır. Tanımlamalar farklı semantik analiz yöntemleri oluşturularak yapılır. Modelleme hedefi ancak aşağıdan yukarıya bir metodoloji ile gerçekleştirilebilir. Alt katmanlarda yer alan özellikler birleştirilerek üst katmanlara erişilebilir.

Yukarıda ifade edilen modelleme fikrini desteklemek üzere dokümanların anlamsal benzerliğini ölçmesi beklenen bir semantik analiz yöntemi önerilmektedir. Yöntem, hipermetin dokümanlarının kümelenmesinde metin

bilgisine ek olarak dokümanlar arasındaki bağ topolojisinin kullanımına dayanır. Manifoldlar üzerindeki jeodezik uzunluk kavramı sayesinde metin ve bağ bilgileri birleştirilir.

Tezde, bilginin manifold üzerinde modellenmesi fikri hakim iken uygulama kısmında önerilen semantik analiz yönteminin de manifold kavramı üzerinden tanımlanması bir rastlantı değildir. Özellikle, bilginin geometrik yapı olarak manifold üzerinde modellenmesi tezini desteklemek için geliştirilmiştir.

Önerilen semantik analiz yaklaşımının ana bileşenlerinin vurgulanması faydalı olabilir:

- Yöntem kapsamında küresel geometri varsayımı ile jeodezik uzunluklar hesaplanmaktadır. Spesifik olarak çember üzerindeki jeodezik uzunluk formülü kullanılmaktadır.

- Önerilen yaklaşımda global bağ çizgesi üzerinde dokümanların yerel bağlantısallık derecelerini ifade eden kümelenme katsayıları, doküman uzayına özgü eğrilik değerlerinin tahminlenmesinde kullanılmaktadır (Lou, 2009).

- Uygulama, k-means kümelenme algoritması ile test edilmektedir.

- Uygulamada k-means kümelenme algoritması sonuçlarını iyileştirmek üzere Cosine metin benzerlik ölçüsü üretilen eğrilik değerleri ile birleştirilir. Bu birleştirme k-means algoritması içinde birden fazla özellik uzayının birleştirilmesi anlamına gelir. Bu bağlamda özellik birleştirme işlevi görecek fonksiyonların test edilebilmesi için algoritma kapsamında tek özellik uzayı ve birden fazla özellik uzaylarının karşılaştırmalı analizinin yapılmasını kapsayacak bir çerçeve tanımlanmasına ihtiyaç vardır. Modha and Spangler (2003) k-means algoritması için özellik ağırlıklandırma bağlamını tanımlayan soyut bir çerçeve ortaya koymuşlardır ve bu tez çalışmasındaki uygulamanın değerlendirme metodolojisini de tanımlarlar.

- Denemeler Wikipedia XML Corpus (Denoyer and Gallinari, 2006) İngilizce altkümesi üzerinde gerçekleştirilmiştir. Değerlendirme

metrikleri kesin referans olarak veri seti ile birlikte gelen Wikipedia kategori bilgisini kullanmaktadır.

- Deneysel sonuçlar, bağ çizgesi üzerinden hesaplanan eğrilik değerlerinin varolan benzerlik ölçülerinin ağırlıklandırılmasında (kümelenme kapsamında kullanılan maliyet fonksiyonunun minimize edilmesinde) kullanılabileceğini ortaya koyar.

5.1 Gelecek Çalışmalar

Deneysel sonuçlar bazı denemelerde jeodezik benzerlik ölçüsünün Cosine'ye göre oldukça iyi olduğunu gösterirken bazı denemelerde aradaki fark çok küçüktür. Araştırılması gereken problemlerden ilki bu durumun nedenini bulmak başka bir deyişle her denemede değişen parametrelerin jeodezik yaklaşımın sonuçlarını nasıl etkilediğini açıklayabilmektir.

Sonuçlar alternatif jeodezik uzunluk hesaplama yöntemlerini teşvik eder. Çalışmada çember üzerindeki jeodezik uzunluk hesaplama formülü kullanılmaktadır. Alternatif olarak uzunluk küre üzerinde hesaplanabilir. Öte yandan, küresel geometri varsayımından tamamen vazgeçilerek, hiperbolik bir uzay üzerinde jeodezik uzunluklar değerlendirilebilir. Zira doküman bağ çizgesinden hesaplanan kümelenme katsayısı değerleri genel olarak 0'a yakındır ve negatif eğrilik eğiliminin görüldüğü bir uzay sözkonusudur. Dolayısıyla bilinen parametreleri kullanacak bir hiperbolik uzaklık hesaplama yöntemi oluşturulabilir.

Ek olarak, kümelenme katsayısı değerlerini eğrilik değerlerine dönüştürmek üzere kullanılan sezgisel algoritma sistemli bir şekilde ele alınıp iyileştirilebilir. Ayrıca jeodezik uzunluk benzerlik ölçüsü başka metin tabanlı bağlı veri setlerinde denenerek sağladığı katkı daha fazla test edilebilir.

“KAYNAKLAR DİZİNİ

Adamic, L.A., Lukose, R.M., Puniyani, A.R. and Huberman, B.A., 2001, Search in power-law networks. *Physical Review E* 64(046135).

Barabási, A.-L., Albert, R. et al., 2000, "Scale-free characteristics of random networks: the topology of the world-wide web." *Physica A: Statistical Mechanics and its Applications* 281(1-4): 69-77.

Batini, C. and Scannapieco, M., 2006, Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), Springer-Verlag New York, Inc, 262p.

Begelfor, E. and Werman, M., 2005, The world is not always flat or learning curved manifolds., School of Engineering and Computer Science, Hebrew University of Jerusalem.

Belkin, M. and Niyogi, P., 2008, "Towards a theoretical foundation for Laplacian-based manifold methods." *J. Comput. Syst. Sci.* 74(8): 1289-1308.

Calado, P., Cristo, M. et al., 2003, Combining link-based and content-based methods for web document classification. Proceedings of the Twelfth International Conference on Information and Knowledge Management. New Orleans, LA, USA, ACM: 394-401.

Cellary, W., 2003, "The profession's role in the global information society." *Computer* 36(9): 124-123.

Clauset, A., Shalizi, C.R., and Newman, M.E.J, 2009, “Power-law distributions in empirical data.” *SIAM Review* 51: 661-703.

Chávez, E., Navarro, G. et al., 2001, "Searching in metric spaces." *ACM Comput. Surv.* 33(3): 273-321.

Cover, T. and Thomas, J., 2006, Elements of Information Theory 2nd Edition, Wiley-Interscience, 776p.

Daconta, M. C., 2003, The Semantic Web : a guide to the future of XML, Web services, and knowledge management / Michael C. Daconta, Leo J. Obrst, Kevin T. Smith. Indianapolis, Ind. Wiley Pub, 312p.

Deerwester, S., Dumais, S. T. et al., 1990, "Indexing by latent semantic analysis." *Journal of the American Society for Information Science* 41(6): 391-407.

KAYNAKLAR DİZİNİ (devam)

Denoyer, L. and Gallinari, P., 2006, "The Wikipedia XML corpus." *SIGIR Forum* 40(1): 64-69.

Dubin, D., 2004, The most influential paper Gerard Salton never wrote. *Library Trends*, Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign. 52**:** 748-764.

Dubrovin, B. A., Fomenko, A. T. et al., 1991, Modern Geometry - Methods and Applications: Part I: The Geometry of Surfaces, Transformation Groups, and Fields (Graduate Texts in Mathematics), Springer, 492p.

Fuhr, N., 2009, Information retrieval models. European Summer School in Information Retrieval, Padova.

Gersh, J., Lewis, B. et al., 2006, "Supporting insight-based information exploration in intelligence analysis." *Commun. ACM* 49(4): 63-68.

Goth, J. and Skrop, A., 2005, "Varying retrieval categoricity using hyperbolic geometry." *Inf. Retr.* 8(2): 265-283.

He, X., Cai, D. et al., 2004, Locality preserving indexing for document representation. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, United Kingdom, ACM: 96-103.

Hui-Fang, M., Qing, H. et al., 2008, Geodesic distance based aproach for sentence similarity computation. Machine Learning and Cybernetics, 2008 International Conference on.

Jonckheere, E. A. and Poonsuk L., 2004, Geometry of network security. American Control Conference (ACC2004), Boston, MA: 976-981.

Jones, K. S., 1988, A statistical interpretation of term specificity and its application in retrieval. *Document retrieval systems*, Taylor Graham Publishing**:** 132-142.

Kamarck, E. C., 2005, Transforming the intelligence community: improving the collection and management of information. Washington, D.C., Report for IBM Center for Business and Government.

KAYNAKLAR DİZİNİ (devam)

Kamps, J. and Koolen, M., 2009, Is Wikipedia link structure different? Proceedings of the Second ACM International Conference on Web Search and Data Mining. Barcelona, Spain, ACM**:** 232-241.

Kantor, P. B. and Lee, J. J., 1986, The maximum entropy principle in information retrieval. Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Palazzo dei Congressi, Pisa, Italy, ACM**:** 269-274.

Kleinberg, J. M., 1999, "Authoritative sources in a hyperlinked environment." *J. ACM* 46(5): 604-632.

Lou, M., 2009, Traffic pattern in negatively curved network. Ph.D. Thesis. University of Southern California, (unpublished).

MacQueen, J. B., 1967, Some methods for classification and analysis of multivariate observations. Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Manning, C., Raghavan, P. et al., 2008, Introduction to Information Retrieval, Cambridge University Press, 498p.

Miller, G. A., 2009, WordNet - About Us. WordNet, Princeton University.

Modha, D. S. and Spangler, W. S., 2003, "Feature weighting in k-means clustering." *Machine Learning* 52: 217-237.

Mukerjee, H. K., 2002, "CAT(κ) Spaces, metric spaces of curvature ≤κ, and sectional curvature of Riemannian manifolds.", www.imsc.res.in/~sankaran/CAT-K.ps (Erişim tarihi: 30 Haziran 2008)

Muller, J. S., 2007, Asymmetry: The Foundation of Information, Springer, 165p.

Newman, M., 2005, "Power laws, Pareto distributions and Zipf's law." *Contemporary Physics* 46(5): 323 - 351.

Oxford English Dictionary, "Social Network. ", http://oxforddictionaries.com/view/entry/m_en_gb0994346#m_en_gb0994346 (Erişim tarihi: 14 Nisan 2010)

Page, L., Brin, S. et al., 1999, The PageRank citation ranking: Bringing order to the Web. Technical Report. Stanford University.

KAYNAKLAR DİZİNİ (devam)

Pal, S., "Java Text Mining Tools.", http://sourceforge.net/projects/jtmt/ (Erişim tarihi: 11 Eylül 2009)

Rand, W., 1971, "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association* 66(336): 846-850.

Ravasz, E., Somera, A. L. et al., 2002, "Hierarchical organization of modularity in metabolic networks." *Science* 297(5586): 1551-1555.

Rijsbergen, C. J. v., 1979, Information Retrieval. London, Butterworths, 208p.

Rijsbergen, C. J. v., 2004, The Geometry of Information Retrieval, Cambridge University Press, 162p.

Riemann, B., 1873, "On the hypotheses which lie at the bases of geometry." *Nature* VIII(183).

Robertson, S. E., 1997, The probability ranking principle in IR. *Readings in Information Retrieval*, Morgan Kaufmann Publishers Inc.**:** 281-286.

Robles-Kelly, A. and Hancock, E. R., 2007, "A Riemannian approach to graph embedding." *Pattern Recognition* 40(3): 1042-1056.

Rowland, T. and Weisstein, E. W., "Geodesic.", http://mathworld.wolfram.com/Geodesic.html. (Erişim tarihi: 27 Ağustos 2008)

Salton, G. and Buckley, C., 1987, Term Weighting Approaches in Automatic Text Retrieval, Cornell University.

Salton, G. and McGill M., 1983, Introduction to Modern Information Retrieval, McGraw-Hill, 400p.

Salton, G., 1989, Automatic text processing : the transformation, analysis, and retrieval of information by computer / Gerard Salton. Reading, Mass. :, Addison-Wesley, 543p.

Searcóid, M. Ó., 2007, Metric Spaces, Springer London.

Shiga, K. and Sunada, T., 2005, A Mathematical Gift, III: The Interplay Between Topology, Functions, Geometry, and Algebra, American Mathematical Society, 129p.

KAYNAKLAR DİZİNİ (devam)

Strehl, A., Strehl, E., et al., 2000, Impact of similarity measures on web-page clustering. In Workshop on Artificial Intelligence for Web Search (AAAI 2000.

Swanson, D. R., 1988, "Historical note: information retrieval and the future of an illusion." *Journal of the American Society for Information Science* 39(2): 94-98.

Ueno, K., Shiga, K. et al., 2003, A Mathematical Gift, 1: The Interplay Between Topology, Functions, Geometry, and Algebra, American Mathematical Society, 136p.

U. S. Congress House Select Committee on Intelligence., 1996, IC21: The Intelligence Community in the 21st Century.

Vinh, N. X., Epps, J., et al., 2009, Information theoretic measures for clusterings comparison: is a correction for chance necessary? Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, ACM**:** 1073-1080.

Waltz, E. L., 1998, Information Warfare Principles and Operations, Artech House, Inc, 416p.

Watts, D. J. and Strogatz, S. H., 1998, "Collective dynamics of /`small-world/' networks." *Nature* 393(6684): 440-442.

White, J. and Kalman, D. Cardano: An Adventure in Algebra in 8 Parts.

Williams, M., 2001, Problems of Knowledge A Critical Introduction to Epistemology, Oxford University Press, 288p.

Wong, S. K. M., Ziarko, W., et al., 1987, "On modeling of information retrieval concepts in vector spaces." *ACM Trans. Database Syst.* 12(2): 299-321.

Wrede, R. C., 1972, Introduction to Vector and Tensor Analysis, Dover Publications, 418p.

Wu, J., H. Xiong, et al., 2009, Adapting the right measures for K-means clustering. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, ACM**:** 877-886.

KAYNAKLAR DİZİNİ (devam)

Xiao, B. and Hancock, E., 2005, “Geometric characterisation of graphs”. Image Analysis and Processing – ICIAP 2005 *LNCS* (3617): 471-478.

Yang, K., 2001, Combining text- and link-based methods for Web IR. Proceedings of the 10th Text Rerieval Conference (TREC-10). Washington, DC., US Government Printing Office.

Zadeh, L. A., 2005, "Toward a generalized theory of uncertainty (GTU): an outline." *Inf. Sci. Inf. Comput. Sci.* 172(1-2): 1-40.

ÖZGEÇMİŞ

Selma Tekir, 30 Haziran 1979 tarihinde Köln'de doğdu. 2001 yılında Ege Üniversitesi Bilgisayar Mühendisliği Bölümü'nden mezun oldu. 2004 yılında İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü'nden yüksek lisansını aldı. 2009 yılında Konstanz Üniversitesi Bilgisayar Bilimleri Fakültesi Veritabanları, Veri Analizi ve Görselleştirme Kürsüsü'nde misafir araştırmacı olarak görev aldı. Halen İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü'nde öğretim görevlisi olarak çalışmaktadır.

Yayınlar

1. Kitap

Tekir S., Open Source Intelligence Analysis A Methodological Approach, VDM Verlag, ISBN: 978-3-639-14036-1, 2009-03-22, 84p.

2. Uluslararası Sempozyum Bildirileri

Tekir, S., "Semantic Data Analysis", First JRC-Turkey Workshop on ICT Security, December 1-2 2009, Yasar University-Izmir, Abstracts of presentations in workshop proceedings.

Koltuksuz A., Tekir, S., "Intelligence Analysis Modeling", International Conference on Hybrid Information Technology, 9-11 Kasım 2006, Cheju Adası, Güney Kore, IEEE Computer Society 10.1109/ICHIT.2006.157, vol. 1, 146-151.

Tekir, S. And Eren Ş., "Information Organization in the Business World", Fifteenth World Business Congress- Technology, Structure, Environment, and Strategy Interfaces in a Changing Global Business Arena, 18-21 Haziran 2006, Saraybosna, Bosna Hersek, International Management Development Research Yearbook, Advances in Global Management Development, vol. XV, 260-263.

Tekir, S., "Open Source Exploitation in Producing Intelligence", 1st International Symposium on Information Technologies, 19-21 Nisan 2005, Girne-KKTC, ISIT 2005 Symposium Proceedings, 84-87.

3. Ulusal Sempozyum Bildirileri

Kurtel, K., Tekir, S., Atay, S., “Lojistik Merkezi Güvenlik Gereksinimleri, XML ve Sayısal İmza”, Ulusal Elektronik İmza Sempozyumu, 7-8 Aralık 2006, Ankara, Bildiriler Kitabı, 116-122.

Tekir, S., Koltuksuz, A., "Bilgi Sistemleri Bilim ve Teknolojisine Dayali Açik Kaynak Istihbarati ve Bir Uygulama Modeli", 2. Polis Bilisim Sempozyumu, 14-15 Nisan 2005, Ankara, Bildiriler Kitabı, 156-161.

Koltuksuz, A., Atay, S., Tapucu, D., Tekir, S., Demiray, O., Asarcıklı, S., Hışıl, H., “Bilgi Sistemleri Güvenligi Dersi”, Akademik Bilisim 2005, 2-4 Subat 2005, Gaziantep Üniversitesi, Bildiri Özetleri Kitabi, 93.