**Федеральное государственное автономное образовательное учреждение   
высшего образования**

**КАЗАНСКИЙ  (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**

**ВЫСШАЯ ШКОЛА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И   
ИНФОРМАЦИОННЫХ СИСТЕМ**

Направление подготовки: 09.03.03 – Прикладная информатика

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**АВТОМАТИЗИРОВАННЫЙ СБОР ДАННЫХ МИРОВЫХ РЕЙТИНГОВ УНИВЕРСИТЕТОВ И ИХ АНАЛИЗ**

**Работа завершена:**

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2017 г.

Студент группы 11-301                             \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Р.М. Аглямов

**Работа допущена к защите:**

Научный руководитель

ученая степень, должность

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2017 г.                   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.А. Марченко

Директор Высшей школы ИТИС

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2017 г.                       \_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.Ф. Хасьянов

Казань – 2017 г.

**Содержание**

[Используемые термины и сокращения 3](#_Toc483982978)

[Введение 4](#_Toc483982979)

[Основная часть 7](#_Toc483982980)

[1. Рассматриваемые рейтинги 7](#_Toc483982981)

[1.1. Academic Ranking of World Universities (ARWU) 7](#_Toc483982982)

[1.2. Quacquarelli Symonds World University Rankings (QS) 11](#_Toc483982983)

[1.3. The Times Higher Education World University Rankings (THE) 16](#_Toc483982984)

[2. Автоматизированный сбор данных 23](#_Toc483982985)

[2.1. Методы 24](#_Toc483982986)

[2.2. Инструменты 24](#_Toc483982987)

[2.3. Этапы 24](#_Toc483982988)

[2.4. Этические сложности 25](#_Toc483982989)

[2.5. Технические сложности 25](#_Toc483982990)

[3. Сбор данных и их анализ 27](#_Toc483982991)

[Заключение 28](#_Toc483982992)

[Список использованных источников 29](#_Toc483982993)

# Используемые термины и сокращения

Веб-скрапинг –

API –

веб-краулинг –

парсинг –

HTML –

Капча –

Веб-мастер –

DOS-атака –

CSV

JSON

RSS

# Введение

Мы живем в эпоху глобализации, затрагивающей буквально все сферы общества. Она не обошла стороной стороной и области образования и научно-исследовательской деятельности. Высокая конкуренция как одна из характерных черт процесса глобализации становится актуальным вызовом для ведущих университетов мира, которые вступают в жесткую конкурентную борьбу. Театром репутационных баталий выступают международно признанные рейтинги ведущих мировых научно-образовательных центров. Место в международных рейтингах стало показателем конкурентного преимущества университетов и своеобразным знаком качества.

Задача получения высокого рейтинга декларируется и ставится на государственном уровне. В «майских указах» 2012 г. президентом В. Путиным была поставлена задача «разработки и утверждения до конца октября 2012 г. плана мероприятий по развитию ведущих университетов, предусматривающих повышение их конкурентоспособности среди ведущих мировых научно-образовательных центров». Контрольным ориентиром государственной политики в области рейтингования образования и науки является необходимость вхождения к 2020 году не менее пяти российских университетов в первую сотню ведущих мировых университетов согласно международным рейтингам.

Работа по исполнению этой и других поставленных задач ведется Министерством образования в рамках проекта «5-100». Проект ориентирован на достижение показателей, удовлетворяющих критериям трех международных глобальных рейтингов:

* The Times Higher Education World University Rankings (THE);
* Academic Ranking of World Universities (ARWU);
* Quacquarelli Symonds World University Rankings (QS).

К сожалению, присутствие российской высшей школы в признанных мировых рейтингах пока немногочисленно. В течение последних 10 лет лишь один российский университет, МГУ им. М.В. Ломоносова, стабильно занимает позиции в пределах 100–300 лучших мировых университетов.

Казанский Федеральный университет (КФУ) также является одним из участников проекта «5-100». На сегодняшний день позиции КФУ не соответствуют тем высоким ожиданиям, которые предполагает данный проект. По итогам рейтинга QS-2016/17  «Казанский Федеральный» занял позицию в группе 401-500 мест. В рейтинге THE-2016/17 вовсе находится за пределами Top-500. Данные результаты заставляют задуматься над причинами столь низких позиций и над тем, что нужно сделать для продвижения. Для выявления причин необходимо осуществить сбор показателей рейтингов и провести анализ. Из-за больших объемов информации, поступающей из разнородных источников и часто обновляющейся, данную работу выполнять вручную неэффективно и крайне ресурсоемко. Необходимо автоматизировать процессы сбора и анализа данных.

[Цель данной дипломной работы осуществить автоматизированный сбор и дальнейший анализ данных показателей рейтингов, на основе которых определить критерии рейтинга, на которые следует сделать основной упор для продвижения в международных рейтингах.]      ???

[Цель данной дипломной работы определить какие критерии международных рейтингов оказывают наибольшее влияние на позиции университетов на основе данных, полученных с помощью автоматизированного сбора, и дать рекомендации для улучшения позиций. Достижение данной цели предполагает решение нескольких задач:] ???

Цель данной дипломной работы разработать рекомендации для улучшения позиций в международных рейтингах КФУ, вузу-участнику проекта «5-100», на основе анализа данных, полученных с помощью автоматизированного сбора. Достижение данной цели предполагает решение нескольких задач:

- проанализировать методологии рейтингов ARWU, QS, THE;

- определить инструменты, с помощью которых будет осуществляться автоматизированный сбор данных;

- исследовать необычные динамики движения позиций университетов в рейтингах и выявить какие критерии оказывали наибольшее влияние на них;

- используя результаты проведенных исследований, разработать и предложить рекомендации по улучшению позиций КФУ в международных рейтингах.

# Основная часть

## Рассматриваемые рейтинги

### 1.1. Academic Ranking of World Universities (ARWU)

В рейтинге ARWU рассматриваются все университеты, в которых есть выпускники-лауреаты или сотрудники-лауреаты Нобелевской или Филдсовской премии, высоко цитируемые исследователи, а также сотрудники, являющиеся авторами статей в журналах Nature и Science. Кроме того, в рейтинг включены университеты, имеющие значительное количество статей, проиндексированных в Science Citation Index-Expanded (SCIE) и Social Science Citation Index (SSCI).

В ARWU ранжируется 1200 университетов, 500 лучших из которых представлены в публикуемой части рейтинга.

**Критерии рейтинга и их веса**

При составлении Академического Рейтинга Университетов Мира ARWU используется 6 индикаторов: число выпускников-лауреатов Нобелевской или Филдсовской премии (Alumni), число сотрудников-лауреатов Нобелевской или Филдсовской премии (Award), число высокоцитируемых исследователей (HiCi), число статей, опубликованных в журналах Nature и Science (N&S), число статей, проиндексированных в Science Citation Index - Expanded и Social Sciences Citation Index (PUB), взвешенный показатель предыдущих индикаторов в расчёте на одного сотрудника (PCP).

Для каждого индикатора университету, получившему наивысшую оценку, присваивается 100 баллов, а показатели остальных университетов рассчитываются как проценты от наивысшего балла. Распределение значений для каждого индикатора изучается на предмет наличия значительных искажений (аномалий), после чего, при необходимости, стандартные статистические методы используются для корректировки индикатора.

Баллы по всем показателям суммируются с весами указанными ниже для получения общей итоговой оценки. Университету, получившему наивысшую оценку, присваивается 100 баллов, а оценки остальных университетов рассчитываются как процент от наивысшего балла. Эти баллы затем размещаются в убывающем порядке.

|  |  |  |  |
| --- | --- | --- | --- |
| Критерий | Индикатор | Код | Вес |
| Качество образования | Число выпускников-лауреатов Нобелевской или Филдсовской премии | Alumni | 10% |
| Квалификация сотрудников | Число сотрудников-лауреатов Нобелевской или Филдсовской премии | Award | 20% |
| Число высоко-цитируемых исследователей в 21 предметной области | HiCi | 20% |
| Достижения в области научных исследований | Число статей, опубликованных в журналах Nature и Science | N&S\* | 20% |
| Число статей, проиндексированных в Science Citation Index - Expanded и Social Sciences Citation Index | PUB | 20% |
| Общая научная производительность на одного сотрудника | Взвешенный показатель предыдущих параметров в расчёте на одного сотрудника | PCP | 10% |
| \* При оценке университетов, специализирующихся на гуманитарных и социальных науках, показатель N&S не учитывается, его доля распределяется между другими показателями. | | | |

Таблица 1 - Индикаторы и их веса для Академического Рейтинга Университетов Мира ARWU

**Определение индикаторов**

**Alumni** – число выпускников-лауреатов Нобелевской и Филдсовской премии. Выпускниками считаются те, кто получил степень бакалавра, магистра или доктора наук в данном университете. Выпускники, окончившие университет в различные периоды времени, учитываются с разными весами. При перемещении на каждые 10 лет назад, вес уменьшается на 10%. Выпускники, окончившие вуз в 2001-2010 гг, имеют вес 100%, в 1991-2000 гг. - 90%, в 1981-1990 гг. - 80%, и т. д., наконец, выпускники 1911-1920 гг. получают вес 10%. Если выпускник университета получил две или более научных степени в данном университете, учитывается только одна.

**Award** – число сотрудников-лауреатов Нобелевской премии (физика, химия, медицина и экономика) и Филдсовской премии (математика). Сотрудниками считаются те, кто работал в университете на момент получения премии. В зависимости от временного промежутка, когда была получена премия, используются различные веса. При перемещении на каждые 10 лет назад, вес  уменьшается на 10%. Сотрудники, получившие премию после 2011 г. получают вес 100%, получившие премию в 2001 -2010 гг. - 90%, в 1991-2000 гг. - 80% и т.д., и, наконец, получившие премию в 1921-1930гг. - 10%. Если лауреат аффилирован более чем с одним университетом, то каждый из них получает вес обратный числу университетов (например, в случае аффилиации с двумя университетами каждый университет получает 50%). Для Нобелевской премии, если премия была получена совместно несколькими учеными, веса распределяются в соответствии с их долей в премии.

**HiCi** - число высоко цитируемых исследователей, выбранных компанией Thomson Reuters. В рейтинге ARWU 2016 года используется список высоко цитируемых исследователей, выпущенный Thomson Reuters 1 декабря 2015 года. Следует отметить, что для тех высоко цитируемых исследователей, кто аффилирован с более чем одним университетом, при расчете индикатора HiCi учитывается только его основное место аффилиации.

**N&S** - Количество статей, опубликованных в журналах Nature и Science в период между 2011 и 2015 годами. При составлении рейтинга учитываются только публикации типа Article. Публикации типа Review или Letter не включаются в расчет рейтинга. Университет, к которому принадлежит ответственный автор статьи, получает вес 100%, университет первого соавтора – 50% (или второго соавтора, если первый соавтор – ответственный автор статьи), университет следующего соавтора – 25%, университеты остальных соавторов по 10% .

**PUB** - Количество статей, проиндексированных в Science Citation IndexExpanded(SCIE) и Social Science Citation Index(SSCI) в 2015 году. При составлении рейтинга учитываются только публикации типа Article. При подсчете общего числа статей университета статьи, индексированные в Social Science Citation Index, имеют удвоенный вес.

**PCP** - общая научная производительность на одного сотрудника: сумма взвешенных баллов вышеупомянутых пяти индикаторов, деленная на число сотрудников (в эквиваленте полной ставки). В случае, когда численность сотрудников университета не доступна, используются взвешенные баллы пяти указанных выше индикаторов.

### 1.2. Quacquarelli Symonds World University Rankings (QS)

Основная цель QS World University Rankings® - помощь студентам при выборе ведущих университетов по всему миру. На основании шести показателей, рейтинг оценивает университеты в четырех областях: исследование, преподавание, трудоустройство и интернационализация.

Каждый из шести показателей имеет различный вес при расчете общего балла университета, его позиции в рейтинге. Четыре показателя основаны на объективных количественных данных, а два - на глобальных опросах ученых и работодателей.

|  |  |
| --- | --- |
| Показатель | Вес |
| Академическая репутация | 40% |
| Репутация среди работодателей | 10% |
| Соотношение научно-педагогического состава и студентов | 20% |
| Количество цитирований на одного сотрудника университета | 20% |
| Доля иностранных преподавателей | 5% |
| Доля иностранных студентов | 5% |

Таблица 2 - Индикаторы и их веса для рейтинга QS

**Академическая репутация (40% от общего балла)**

Академическая репутация рассчитывается с помощью глобального опроса в научной среде. Ученые должны указать университеты, которые, по их мнению, выделяются с точки зрения проведения научной работы в их области знаний.

Для рейтинга 2016/17 года получены ответы 74651 ученого по всему миру. Принимающие участие в опросе ученые не могут голосовать за свой университет. При анализе данных применяются различные региональные коэффициенты с целью сглаживания страновых различий.

При составлении рейтинга эксперты разделяются на «международных» и «местных». Оценки «международных» и «местных» экспертов масштабируются так, чтобы лучший университет имел оценку 100  (создается две отдельные шкалы). Две оценки складываются с весами 85% для «международных» экспертов и 15% - для «местных». Из полученной суммы извлекается квадратный корень с целью сокращения разрыва между самыми сильными и самыми слабыми университетами, чтобы превосходство в одном из пяти отраслевых направлений (Faculty) не давало чрезмерных преимуществ при ранжировании в общем рейтинге.

Затем полученные результаты масштабируются так, чтобы лучший университет в определенном отраслевом направлении имел оценку 100. Далее для каждого университета оценки, полученные по различным отраслевым направлениям, складываются с равными весами. Полученная оценка стандартизируется на основе выборки университетов.

Для расчёта показателя используются данные за пять последних лет. При этом до 2015 года в рейтинге учитывался только последний ответ конкретного респондента за последние три года, с 2015 года также учитываются данные, полученные 4-5 лет назад (также только последний ответ респондента). Однако полученные таким образом результаты умножаются на понижающий коэффициент (0,5 и 0,25, соответственно) для их сопоставления с данными за последние три года.

**Репутация среди работодателей (10%)**

Репутация среди работодателей также рассчитывается с помощью опроса. Для рейтинга 2016/17 был опрошен 37781 респондент – работодатель. В рамках опроса работодателей просят определить университеты с лучшими, наиболее конкурентоспособными выпускниками. Этот показатель является уникальным среди международных университетских рейтингов.

Чтобы компенсировать разницу между странами с большим и малым количеством предоставленных анкет, используются специальные коэффициенты.

Больший вес имеют ответы зарубежных работодателей по отношению к оцениваемому университету, так как это выделяет университеты, имеющие международную репутацию.

При составлении рейтинга эксперты разделяются на «международных» и «местных». Оценки «международных» и «местных» экспертов масштабируются так, чтобы лучший университет имел оценку 100 (создается две отдельные шкалы). Две оценки складываются с весами 70% для «международных» экспертов и 30% - для «местных». Из полученной суммы извлекается квадратный корень с целью сокращения разрыва между самыми сильными и самыми слабыми университетами, чтобы превосходство в одном из пяти отраслевых направлений не давало чрезмерных преимуществ при ранжировании в общем рейтинге.

Затем полученные результаты масштабируются так, чтобы лучший университет в определенном отраслевом направлении имел оценку 100. Далее для каждого университета оценки, полученные по различным отраслевым направлениям, складываются с равными весами. Полученная оценка стандартизируется на основе выборки университетов.

Для расчёта показателя используются данные за пять последних лет. При этом до 2015 года в рейтинге учитывался только последний ответ конкретного респондента за последние три года, с 2015 года также учитываются данные, полученные 4-5 лет назад (также только последний ответ респондента). Однако полученные таким образом результаты умножаются на понижающий коэффициент (0,5 и 0,25, соответственно) для их сопоставления с данными за последние три года.

**Соотношение научно-педагогического состава и студентов (20%)**

Соотношение научно-педагогического состава и студентов рассчитывается как отношение количества сотрудников университета из расчета полной ставки (здесь и далее: работающие на полную ставку – 1, работающие на неполную ставку – 1/3) к количеству студентов (Full Time Equivalent, некоторый аналог приведенного контингента). При расчете рейтинга под студентами понимаются как непосредственно студенты, так и аспиранты.

При отсутствии международного стандарта, на основании которого можно измерить качество обучения, этот показатель направлен на выявление университетов, предоставляющих наилучшие условия для студентов как с точки зрения возможности обучения небольшими группами, так и с точки зрения индивидуального подхода к обучению.

**Количество цитирований на одного сотрудника университета (20%)**

Количество цитирований на одного сотрудника университета рассчитывается, как отношение количества цитирований на работы сотрудников вуза за последние пять полных лет к общему количеству сотрудников из расчета полной ставки (таким образом, размер университета не влияет на данный показатель). Показатель рассчитывается на основании базы данных Scopus.

Данный показатель направлен на оценку научного влияния университетов. Цитирование означает признание проведенного исследования другим исследователем. Чем больше цитирований получает исследование, тем больше влияния оно имеет в мире. Чем больше в университете высоко цитируемых работ, тем выше его научное влияние и репутация.

С 2015 года при составлении рейтинга производится нормализация индекса цитирования: до 2015 года при анализе данных Scopus для QS World University Rankings рассматривались все результаты в целом, однако в связи с неравномерным распределением продуктивности ученых в каждой из областей знаний было принято решение отобразить это в анализе данных. Процесс «нормализации» (корректировки) заключается в оценке индекса цитируемости, не в целом, а по конкретным группам наук: искусство и гуманитарные науки; общественные науки, включая управление; естественные науки; технические науки и инженерное дело; медико биологические науки; помимо этого, в двух областях - искусство и гуманитарные науки, а также общественные науки, включая управление - были введены дополнительные корректирующие коэффициенты для различных стран. Таким образом, была разработана скользящая шкала, позволяющая оценить и нормализовать индекс цитирования в зависимости не только от общего количества цитат, но и от языка публикации, продуктивности страны в той или иной области знаний в целом и т.д.

Помимо этого, из анализа исключаются публикации с количеством аффилиаций авторов больше определенного заданного количества в зависимости от предметной области.

Для составления рейтинга 2016/17 года было проанализировано 10,3 млн научных статей и 66,3 млн цитирований. Самоцитирования из анализа исключаются (в связи с этим в итоговом анализе из 66,3 млн цитирований учитывалось только 50,4 млн цитирований).

**Доля иностранных сотрудников (5%)**

Доля иностранных сотрудников рассчитывается как отношение количества иностранных сотрудников к общему количеству сотрудников университета. В расчете учитываются иностранные сотрудники, которые проработали в университете не менее трех месяцев за анализируемый год.

**Доля иностранных студентов (5%)**

Доля иностранных студентов рассчитывается как отношение количества иностранных студентов к общему количеству студентов в университете (Full Time Equivalent).

Цель последних двух показателей - оценка степени успешности университета в привлечении ученых и студентов из других стран.

### 

### 1.3. The Times Higher Education World University Rankings (THE)

Рейтинг мировых университетов Times Higher Education является единственным рейтингом, который оценивает университеты, активно проводящие исследования, по всем их направлениям деятельности: преподавание, исследования, передача знаний и международное взаимодействие. При составлении рейтинга используются 13 индикаторов эффективности с целью обеспечения наиболее полного и сбалансированного сравнения, которому доверяют студенты, ученые, руководители университетов и промышленности и правительства. Индикаторы эффективности сгруппированы по пяти направлениям:

1. Преподавание (среда обучения).

2. Исследования (объем, доход и репутация).

3. Цитирования (влияние исследований).

4. Международное взаимодействие (сотрудники, студенты и исследования).

5. Доход от производственной деятельности (передача знаний).

В 2016 году впервые в публикуемую часть рейтинга вошло 980 университетов.

|  |  |  |
| --- | --- | --- |
| Критерий | Индикатор | Вес |
| Преподавание (среда обучения) | Академическая репутация в сфере образования | 15% |
| Отношение числа преподавателей к числу студентов | 4,5% |
| Отношение выпускников с PhD и выпускников-бакалавров | 2,25% |
| Отношение числа присужденных докторских степеней к числу сотрудников университета | 6% |
| Доход института | 2,25% |
| Исследования (объем, доход и репутация) | Исследовательская репутация | 18% |
| Доход от исследовательской деятельности | 6% |
| Продуктивность исследований | 6% |
| Цитирования (влияние исследований) |  | 30% |
| Международное взаимодействие (сотрудники, студенты и исследования) | Отношение числа иностранных и местных студентов | 2,5% |
| Отношение числа иностранных и местных сотрудников университета | 2,5% |
| Международное сотрудничество (Доля научных публикаций, написанных в соавторстве с иностранными учёными) | 2,5% |
| Доход от производственной деятельности (инноваций) (передача знаний): |  | 2,5% |

Таблица 3 - Индикаторы и их веса для рейтинга THE

1. **Преподавание (среда обучения): 30%**

**1. Академическая репутация в сфере образования (Репутационный опрос): 15%**

Опрос экспертов, выявляющий академическую репутацию университетов в преподавании (проводится ежегодно), проводился с января 3 по март 2016 года. Ответы были статистически репрезентативны по географической и предметной принадлежности ученых. В опросе приняли участие 10323 эксперта. Данные 2016 года были объединены с данными опроса 2015 года. В результате при составлении рейтинга 2016 года анализировались ответы более чем 20000 экспертов.

1. **Отношение числа преподавателей к числу студентов: 4,5%**

**3. Отношение выпускников с PhD и выпускников-бакалавров: 2,25%**

**4. Отношение числа присужденных докторских степеней к числу сотрудников университета: 6%.**

Данный индикатор позволяет понять, во-первых, насколько университет посвящает себя воспитанию следующего поколения ученых, во- вторых, какова в университете доля аспирантов, так как большое количество аспирантов является показателем осуществления преподавания на высочайшем уровне (этот факт является привлекательным для студентов). Этот индикатор нормирован так, чтобы учесть уникальный набор предметов для данного университета, так как количество присужденных докторских степеней варьируется от дисциплины к дисциплине.

**5. Доход института: 2,25%**

Данный индикатор учитывает число сотрудников университета и нормализован по паритету покупательной способности. Подобный подход наиболее наглядно показывает общий статус университета, а также развитость его инфраструктуры и наличие оборудования, доступного для студентов и сотрудников.

1. **Исследования (объем, доход и репутация): 30%**

**1. Исследовательская репутация (Репутационный опрос): 18%**

Наиболее важный индикатор в категории «Исследования» указывает на репутацию университета в сфере исследований на основе проводимого ежегодного опроса академической репутации экспертов. При этом для анализа репутационной составляющей университета в сфере исследований учитываются ответы экспертов, представляющих ту же область, что и университет.

**2. Доход от исследовательской деятельности: 6%**

Данный индикатор учитывает число сотрудников, работающих в университете, и нормируется по паритету покупательной способности. Этот индикатор может быть подвержен влиянию национальной политики и экономических обстоятельств, но доход является важным для развития исследований мирового класса, и поскольку большая его часть является результатом конкурсной борьбы, эксперты, участвующие в составлении рейтинга, полагают, что данный индикатор важен для анализа. Этот индикатор нормирован по предметным профилям каждого университета, так как размер выделяемых грантов сильно зависит от научной области (часто гранты для социальных наук, искусства и гуманитарных наук меньше, чем в других научных областях).

**3. Продуктивность исследований: 6%**

В данном индикаторе анализируется количество статей на одного ученого, опубликованных в научных журналах индексируемых базой данных Scopus. Данный индикатор учитывает размер института и нормализован по предметной области.

1. **Цитирования (влияние исследований): 30%**

Данный индикатор уделяет внимание роли университета в распространении новых знаний и идей.

В данном индикаторе рассматривается влияние исследований на основе расчета количества цитирований, которое работа, опубликованная университетом, получила по всему миру. При составлении рейтинга 2016- 2017 годов поставщик библиометрических данных Elsevier проверил более 56 миллионов цитирований на 11,9 миллиона журнальных статей, докладов конференций, книг и глав книг, опубликованных за пять лет (до 2015 года 5 поставщиком библиометрической информации (публикации, цитируемость) являлась компания Thomson Reuters и, соответственно, использовалась база данных Web of Science; при составлении рейтинга 2015-2016 года в качестве поставщика библиометрических данных стала использоваться компания Elsevier и, соответственно, база данных Scopus). Эти данные извлечены из 23000 научных журналов, проиндексированных базой данных Scopus, и включают все проиндексированные журналы, опубликованные между 2011 и 2015 годами.

Цитирования этих статей учитываются за шестилетний период: с 2011 по 2016 год. Цитирования помогают понять вклад каждого университета в сумму человеческих знаний: они демонстрируют, чьи исследования являются выдающимися, были использованы и востребованы другими учеными и, что наиболее важно, были переданы всему мировому научному сообществу, независимо от научной дисциплины.

Данные полностью нормированы, чтобы отражать вариации в количестве цитирований между различными предметными областями. Это означает, что университеты с высоким уровнем исследовательской активности по предметам с традиционно высоким уровнем цитирования не получают нечестного преимущества.

При составлении рейтинга данные по цитированию отчасти нормируются по странам с целью уменьшения влияния измерения числа цитирований англоязычных публикаций: нормированный и ненормированный по странам уровни цитирований комбинируются с равными весами.

При составлении рейтинга 2015-2016 года из индикатора цитирования были удалены публикации с более чем 1000 авторами (всего - 649 статей). При составлении рейтинга 2016-2017 года подобные статьи были включены в анализ, но использовалась определенная система весов: университет, сотрудники которого являются соавторами таких работ, получает не менее 5% (5% - за каждого автора).

1. **Международное взаимодействие (сотрудники, студенты и исследования): 7,5%**

**1. Отношение числа иностранных и местных студентов: 2,5%**

**2. Отношение числа иностранных и местных сотрудников университета: 2,5%**

Способность университета привлекать студентов, аспирантов и сотрудников со всей планеты – это ключ к успеху на мировой арене.

**3. Международное сотрудничество (Доля научных публикаций, написанных в соавторстве с иностранными учёными): 2,5%**

Данный индикатор рассчитывается как отношение количества научных работ, написанных в соавторстве по крайней мере с одним иностранным автором к общему количеству научных публикаций вуза. Индикатор нормализован по набору предметных областей университета и учитывает тот же пятилетний период публикаций, что и в категории «Цитирования: влияние исследований».

1. **Доход от производственной деятельности (инноваций) (передача знаний): 2,5%**

Способность университета помогать промышленности инновациями, изобретениями и консультациями стало центральной задачей современной научной организации. Данный индикатор основывается на уровне дохода университета от индустрии (нормированный на паритет покупательной способности) и учитывает количество ученых, которые в нем работает.

Этот индикатор демонстрирует степень желания бизнеса платить за исследования, проводимые университетом, а также способность университета привлекать финансирование на коммерческом рынке.

## Автоматизированный сбор данных

Автоматизированный сбор данных из Интернета существует столько же, сколько сам Интернет. То, что мы видим как веб-сайт с отзывом о фильме, для машины представляется сборищем разных «кусков» данных с непонятным назначением. Человек, взглянув на веб-страницу, сразу легко определяет нужный и значимый раздел, но для компьютера понимание того, какой именно текст следует обрабатывать, как отделить этот текст от рекламы, ненужных заголовков, ссылок является довольно сложной задачей.

По мере роста потока информации, возможностей по применению этой информации в прикладных задачах, развиваются технические подходы, объединяемые общим термином «веб-краулинг» или «веб-скрапинг». Они предназначены для сбора информации из сети Интернет и ее подготовки к автоматизированной обработке.

Несмотря на то, что веб-скрапинг (web scraping) не является новым термином, раньше это направление было больше известно под названием анализ экранных или интерфейсных данных (screen scraping), интеллектуальный анализ данных (data mining), сбор веб-данных (web harvesting). Похоже, что на сегодняшний день общее мнение склоняется в пользу термина веб-скрапинг (web scraping), который и будет использоваться в рамках данной работы.

В теории веб-cкрапинг – это сбор данных с помощью любых средств, кроме программ, использующих API. Чаще всего веб-скрапинг осуществляется с помощью программы, которая автоматически запрашивает веб-сервер, запрашивает данные (HTML и другие файлы, которые размещены на веб-страницах), а затем выполняет парсинг этих данных, чтобы извлечь необходимую информацию.

На практике веб-скрапинг охватывает широкий спектр методов и технологий программирования, таких как анализ данных и информационная безопасность.

### Методы

выфвыфвыф

### Инструменты

### Этапы

Парсинг HTML-страницы представляет из себя процесс, который можно разбить на три этапа:

1. **Получение исходного кода веб-страницы.** В разных языках для этого предусмотрены различные способы. Например, в PHP чаще всего используют библиотеку cURL или же встроенную функцию file\_get\_contents.
2. **Извлечение из html-кода необходимых данных.** Получив страницу, необходимо обработать её – отделить обычный текст от гипертекстовой разметки, выстроить иерархическое дерево элементов документа, корректно среагировать на невалидный код, извлечь со страницы необходимую информацию.
3. **Фиксация результата.** Благополучно обработав данные на странице, требуется их сохранить в необходимом виде для последующей обработки. Спарсенное обычно заносится в базу данных, однако есть и другие варианты. Иногда требуется записать в CSV-файл или строить иерархические JSON-структуры, иногда сконвертировать в excel-таблицу, а может даже сгенерировать динамический RSS-поток.

Как правило, требуется спарсить не одну страницу сайта-донора, а множество, может даже и все. В этом случае после прохождении шагов 1-3 в алгоритм парсера должен быть заложен переход на следующую страницу сайта, дабы и с неё извлечь необходимый материал.

Обход всех нужных страниц сайта обеспечивается разными способами.

**Во-первых,** обрабатывая очередную страницу, парсер можно научить не только извлекать необходимые данные, но и заносить в свою базу данных все внутренние ссылки, встречающиеся по пути. Обращаясь к своему хранилищу ссылок, программа последовательно посещает страницы сайта, до тех пор пока не обойдёт их все.

**Во-вторых,** при первичном анализе сайта зачастую возможно проследить логику формирования url для страниц. И затем, генерировать адреса в соответствии с выявленными закономерностями.

**В-третьих,** некоторые парсеры рассчитаны, как ни странно, на «ручной» обход веб-ресурса. Пользователь, кликая по ссылкам, сам решает какие страницы посещать, какие нет. А программа в фоновом режиме запоминает необходимые данные.

Разумеется, совмещать разные методы ничто не мешает.[[1]](#footnote-1)

### Этические сложности

Вопрос о том, является ли парсинг воровством контента, активно обсуждается во Всемирной сети. Большинство оппонентов считают, что заимствование части контента, не являющегося интеллектуальной собственностью, например, технических описаний, допустимо. Ссылка на первоисточник контента рассматривается как способ частичной легитимации. В то же время, наглое копирование, включая грамматические ошибки, осуждается интернет-сообществом, а поисковыми системами рассматривается как повод для блокировки ресурса.[[2]](#footnote-2)

### Технические сложности

Кроме этических проблем парсер способен создать и технические. Он автомат, робот, но его вход на сайт фиксируется, а входящий и исходящий трафики учитываются. Количество подключений к веб-ресурсу в секунду устанавливает создатель программы. Делать этот параметр очень большим нельзя, поскольку сервер может не выдержать количество потока входящего трафика. При частоте 200–250 подключений в секунду работа парсера рассматривается как аналогичная *DOS-атаке*. В связи с этим веб-мастера могут применять различные меры борьбы с автоматическими парсерами, а именно:

1. Динамическое изменение структуры кода страницы с целью усложнения извлечения информации из блоков.
2. Показ капчи при авторизации/регистрации/превышении числа запросов за единицу времени.
3. Блокировка клиентов по среднему объему трафика в единицу времени.
4. Анализ поведения клиентов и блокировка/требование пройти капчу для продолжения работы при подозрительном поведении.

## 3. Сбор данных и их анализ

# Заключение

# Список использованных источников

1. Этапы парсинга – Парсинг: Что? Зачем? Как? [Электронный ресурс]. – Режим доступа: http://parsing.valemak.com/ru/what-why-how/stages-of-parsing/. – (Дата обращения: 09.03.2017). [↑](#footnote-ref-1)
2. Парсинг. Что это и где используется – iPipe [Электронный ресурс]. – Режим доступа: https://www.ipipe.ru/info/parsing.html/. – (Дата обращения: 09.03.2017). [↑](#footnote-ref-2)